RECOVERING RANK AXES IN EMBEDDING SPACES VIA DIFFERENTIABLE RANKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent studies have shown that visual embeddings contain latent directions along which projections preserve ordinal attributes such as age, crowd count, and aesthetic quality. These *rank axes* provide an interpretable way to uncover ordinal structure in pretrained representations, yet existing estimation methods do not directly target ordinal agreement. In this paper, we study rank axis estimation in a fixed embedding space through differentiable ranking and ask when minimizing a rank-based objective recovers the underlying ordinal direction. We consider two estimators based on quadratically regularized ranking over the permutahedron, using a Spearman-type squared rank loss and a Fenchel–Young loss. We establish consistency of the Spearman-based estimator under Gaussian noise and of the Fenchel–Young estimator in the noise-free setting. Notably, their behavior under noise differs, showing that closely related differentiable ranking objectives can have different recovery properties. We also establish convergence to first-order stationarity for both optimization procedures. Experiments on synthetic and real-world image datasets demonstrate accurate rank-axis recovery and improved overall and top-k ranking performance over linear regression and learning-to-rank baselines. Our results show that differentiable ranking provides an effective framework for recovering ordinal structure in pretrained embeddings, while highlighting the statistical consequences of the choice of ranking objective.

1 INTRODUCTION

Modern machine learning systems increasingly rely on high-dimensional embeddings to represent images, text, and other complex data. A central goal of representation learning is to organize these embeddings so that semantic relationships are reflected in their geometry. Most approaches emphasize similarity, encouraging semantically similar inputs to lie close together and encoding meaning through proximity (Chen et al., 2020; Radford et al., 2021). Yet similarity is only one form of structure in an embedding space. Learned representations may also encode ordinal information through directions along which objects can be meaningfully ordered.

A growing body of work suggests that semantic attributes often exhibit approximately linear structure in learned representations (Park et al., 2024; Zou et al., 2023). More recently, this observation has been extended to continuous attributes with an inherent ordinal structure. Attributes such as age, crowd count, aesthetic quality, and temporal recency can often be recovered by projecting embeddings onto a suitable direction and sorting the resulting scalar projections (Sonthalia et al., 2025). Such directions, referred to as *rank axes*, provide a simple and interpretable mechanism for turning a high-dimensional embedding space into an attribute-aware one-dimensional ordering, with potential applications to attribute-aware retrieval and ranking.

The key question is therefore not merely whether an attribute can be predicted from an embedding, but whether the embedding contains a direction that preserves its ordering. This distinction is important because existing methods for recovering rank axes do not directly target this objective. A common approach fits a linear regression of the attribute values on the embeddings (Sonthalia et al., 2025). Regression, however, depends on the numerical magnitudes and spacing of the attribute values even when only their ordering is of interest. Another approach constructs a direction from differences between samples at the extremes of the attribute range (Sonthalia et al., 2025), thereby leaving much of the available ordering information unused. Earlier work on relative and ordinal at-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

tributes developed ranking-based approaches for visual attributes (Parikh & Grauman, 2011), while more recent work has adapted large pretrained vision-language models for ordinal classification (Wang et al., 2023) and image ranking (Yu et al., 2025). In contrast, our focus is on recovering a single global ordinal direction already present in a fixed pretrained embedding space.

A natural approach is to estimate the rank axis by directly optimizing agreement between the projected and target orderings. Differentiable ranking makes this possible by replacing discrete ranks with soft approximations. In particular, regularized ranking over the permutahedron provides a rank operator that is differentiable almost everywhere, enabling Spearman-type criteria (Blondel et al., 2020b), while Fenchel–Young losses provide a general framework for regularized prediction and are differentiable everywhere (Blondel et al., 2020a). This raises a fundamental statistical question. *Does minimizing a differentiable ranking loss recover the underlying rank axis?* Such recovery depends critically on the choice of loss, as objectives based on related soft-ranking operators may have different population minimizers.

We investigate this question using two objectives based on quadratically regularized ranking over the permutahedron: a Spearman-type squared rank loss and a Fenchel–Young loss. We characterize when these objectives recover the underlying rank axis and how their statistical recovery properties differ. Our theoretical results characterize this distinction from both statistical and optimization perspectives. For the Spearman-based estimator, we establish consistency under a Gaussian noise model, showing that minimizing the empirical rank-based criterion asymptotically recovers the underlying rank direction under the stated conditions. For the Fenchel–Young estimator, we establish consistency in the noise-free setting. Notably, the behavior under noise differs, demonstrating that losses based on closely related differentiable ranking constructions need not share the same recovery properties. These results show that differentiability and ranking awareness alone are insufficient to guarantee statistical recovery. Separately, we establish convergence to first-order stationarity for the gradient-based optimization procedures associated with both objectives. These optimization guarantees concern the computation of stationary solutions and are distinct from the statistical consistency results, which characterize global minimizers of the corresponding population objectives.

Our main contributions are as follows. (i) We study rank axis estimation from the perspective of statistical recovery. Building on the notion of rankability in embedding spaces (Sonthalia et al., 2025), we consider the recovery of a global ordinal direction in a fixed embedding space and ask when differentiable rank-based objectives recover the underlying direction. We analyze two estimators based on quadratically regularized ranking over the permutahedron, using a Spearman-type squared rank loss and a Fenchel–Young loss (Section 3.1). (ii) We establish convergence to first-order stationarity for the optimization procedures associated with both objectives, distinguishing these computational guarantees from the statistical recovery guarantees for global minimizers (Section 3.2). (iii) We establish consistency of the Spearman-based estimator under Gaussian noise and of the Fenchel–Young estimator in the noise-free setting, while showing that closely related differentiable ranking objectives can exhibit different recovery behavior under noise (Section 3.3). (iv) We investigate rank-axis recovery on synthetic data and evaluate the methods across multiple real-world datasets and pretrained embedding models, demonstrating improvements in both overall and top-k ranking performance over existing baselines (Section 4).

Taken together, our results shift the focus from whether differentiable ranking can be used to estimate a rank axis to when the resulting estimator recovers the underlying ordinal direction. They show that ranking-aware objectives can uncover ordinal structure in pretrained embeddings, while closely related soft-ranking objectives may have fundamentally different recovery properties.

2 Related Works

Ordinal Structures in Embedding Spaces Prior work has used linear methods to probe and interpret semantic structure in learned representations, including linear classifiers and concept directions (Mikolov et al., 2013; Alain & Bengio, 2017; Kim et al., 2018; Park et al., 2024; Marks & Tegmark, 2024). Continuous attributes reveal a different form of linear structure, in which a projection direction is expected to preserve the relative ordering of samples rather than merely separate semantic categories. Closely related to this idea, Parikh & Grauman (2011) introduce linear ranking functions to encode relative visual attributes, with subsequent work extending relative-attribute learning to deep visual representations (Singh & Lee, 2016). More recently, Sonthalia et al. (2025) introduce

the notion of rankability in visual embeddings and show that continuous attributes can be represented along linear rank axes. They consider several strategies for recovering such axes, including linear regression, few-shot methods, and zero-shot text-based directions in vision-language models. Nam et al. (2026) further study zero-shot rankability in multimodal embeddings by constructing ordinal directions from language. Related work has also adapted vision language models explicitly for ordinal prediction and image ranking through ordering-aware prompt learning and adapter-based methods (Wang et al., 2023; Yu et al., 2025). Against this background, we focus on rank axis estimation in fixed pretrained embedding spaces. Our main distinction lies in the estimation objective: although existing rank axis methods are evaluated by rank correlation, their estimation procedures do not directly optimize ordinal agreement. We build on the notion of rank axes by making this agreement the estimation objective itself, and develop a ranking-based framework with both statistical and optimization guarantees.

Ordinal Regression and Learning to Rank Learning from ordered targets has been widely studied in ordinal regression and learning to rank. Ordinal regression methods explicitly exploit the ordering among labels, with threshold-based formulations providing a general framework for ordinal prediction (Crammer & Singer, 2001). Such methods have been applied to tasks including age estimation (Niu et al., 2016; Cao et al., 2020) and depth estimation (Fu et al., 2018). Learning-to-rank methods are commonly categorized into pointwise, pairwise, and listwise (Liu, 2011) approaches. Pairwise methods, such as RankNet (Burges et al., 2005), Ranking SVM (Herbrich et al., 2000), and RankBoost (Freund et al., 2003), learn from pairwise preferences, while listwise methods define objectives directly over ranked lists. For example, ListNet (Cao et al., 2007) minimizes the cross-entropy between distributions induced by predicted and target rankings, whereas ListMLE (Xia et al., 2008) formulates ranking through a likelihood over permutations. Ranking information has also been incorporated directly into representation learning. Rank-N-Contrast (Zha et al., 2023), for instance, learns continuous representations that reflect the ordering of target attributes. In contrast to general learning-to-rank methods, we seek to recover a single interpretable direction in a fixed modern embedding space whose projection preserves the global ordering of a target attribute.

Differentiable Ranking Direct optimization of rank-based objectives is challenging because sorting and ranking are inherently non-differentiable. Early approaches such as SoftRank (Taylor et al., 2008) introduce smooth approximations to ranking metrics, while SoDeep (Engilberge et al., 2019) learns differentiable surrogates for sorting and applies them to ranking losses, including Spearman rank correlation. Other work develops continuous relaxations of sorting through permutation matrices (Grover et al., 2019), optimal transport (Cuturi et al., 2019), or relaxations of the `argsort` operator (Prillo & Eisenschlos, 2020). An alternative approach by Blondel et al. (2020b) formulates sorting and ranking as convex optimization problems over the permutahedron, yielding efficient differentiable operators computed using isotonic regression, including the pool adjacent violators algorithm (PAVA) (Barlow & Brunk, 1972). Building on this framework, Sander et al. (2023) develop soft top-k operators using p-norm regularization, while Blondel et al. (2020a) establish the corresponding Fenchel–Young loss formulation. We use the quadratically regularized ($p = 2$) soft rank and soft top-k operators to construct differentiable ordinal objectives for rank axis estimation, together with the corresponding Fenchel–Young loss.

3 Proposed methods

A rank axis is intended to capture the ordering of an attribute rather than its numerical scale. Consequently, a regression-based axis estimator may depend on the magnitudes and spacing of the attribute values even when different value assignments induce exactly the same ordering. This motivates estimating the axis by directly maximizing agreement between the ordering induced by projected embeddings and the target ordering of the attribute.

Let $\{(f_i, a_i)\}_{i=1}^n$ be a dataset of n samples, where $f_i \in \mathbb{R}^d$ is the embedding vector of the i-th image and $a_i \in \mathbb{R}$ is a scalar attribute (e.g., age, count). Our goal is to find a unit vector $v \in \mathbb{S}^{d-1}$ such that the ordering of the projected scores $s_i := v^\top f_i$ agrees with the ordering of the attribute values a_i.

3.1 Rank Axis Learning via Differentiable Ranking

To enable direct optimization of rank-based objectives, we use a differentiable approximation to the hard rank operator. The hard rank operator is piecewise constant, with zero gradient almost everywhere, and changes discontinuously when the ordering of the scores changes. A soft rank operator replaces these discrete ranks with smooth values, providing informative gradients while retaining the underlying ordering structure.

Define the soft rank $r_n(s(v))$ of the projected scores via regularized projection onto the permutahedron (Blondel et al., 2020b):

$$r_n(s(v)) := \underset{y \in \mathcal{P}(n)}{\arg\min} \left\{ \langle -s(v), y \rangle + \frac{l_n}{2} \|y\|_2^2 \right\}, \tag{1}$$

where $s(v) = (v^\top f_1, \ldots, v^\top f_n) \in \mathbb{R}^n$, $\mathcal{P}(n)$ is the permutahedron generated by all permutations of $\{1, \ldots, n\}$, and $l_n > 0$ is the regularization parameter for soft rank. For distinct scores, the soft rank approaches the corresponding hard rank as $l_n \to 0$. The ground-truth rank vector is $t^{(n)} := \mathrm{rank}([a_1, \ldots, a_n])$ with average ranks assigned to ties.

We consider the following two objectives for learning the rank axis. The first is the Spearman-type squared rank loss,

$$L_n^{\mathrm{Sp}}(v) := \frac{1}{n^3} \left\| r_n(s(v)) - t^{(n)} \right\|_2^2. \tag{2}$$

The factor n^{-3} normalizes the squared rank discrepancy, whose natural scale is of order n^3. The second is the Fenchel–Young loss (Blondel et al., 2020a),

$$L_n^{\mathrm{FY}}(v) := \frac{1}{n} \sup_{y \in \mathcal{P}(n)} \left(\langle s(v), y \rangle - \frac{l_n}{2} \|y\|_2^2 + \frac{l_n}{2} \left\| t^{(n)} \right\|_2^2 - \left\langle s(v), t^{(n)} \right\rangle \right). \tag{3}$$

The Spearman-type squared rank loss directly measures the discrepancy between the predicted soft ranks and the target ranks, and is therefore closely aligned with rank-based evaluation. In contrast, the Fenchel–Young loss is induced by the regularized ranking operator, and its gradient with respect to the scores is proportional to the difference between the predicted soft ranks and the target ranks. For brevity, we hereafter refer to this Spearman-type squared rank loss as the Spearman loss.

The soft rank operator is differentiable almost everywhere with respect to its score (Blondel et al., 2020b). At points where the loss is differentiable, $\nabla_v L(v)$ denotes the ordinary Euclidean gradient. At nondifferentiable points of the Spearman loss, we use $\nabla_v L(v)$ to denote a limiting gradient from an adjacent differentiable region, with a slight abuse of notation. We optimize either objective on the unit sphere using the projected update direction $g = (I - vv^\top)\nabla_v L(v)$, as summarized in Algorithm 1.

Algorithm 1 Learning rank axis via differentiable ranking (permutahedron projection)

Require: $\{(f_i, a_i)\}_{i=1}^n$, $l_n > 0$, learning rates $\{\eta_k\}_{k \geq 0}$, loss $L \in \{L_n^{\mathrm{Sp}}, L_n^{\mathrm{FY}}\}$
Ensure: Rank axis $v \in \mathbb{R}^d$ with $\|v\|_2 = 1$

$t^{(n)} \leftarrow \mathrm{rank}([a_1, \ldots, a_n])$ ▷ ground-truth rank vector
Initialize $v_0 \in \mathbb{S}^{d-1}$
for $k = 0, 1, \ldots$ until convergence **do**
  $s(v_k) \leftarrow [v_k^\top f_1, \ldots, v_k^\top f_n]$ ▷ projected scores
  $r_n(s(v_k)) \leftarrow \arg\min_{y \in \mathcal{P}(n)} \left\{ \langle -s(v_k), y \rangle + l_n \|y\|_2^2 / 2 \right\}$ ▷ soft rank vector
  Compute $\nabla_v L(v_k)$ by backpropagation through the projection
  $g_k \leftarrow (I - v_k v_k^\top)\nabla_v L(v_k)$ ▷ projected update direction
  $v_{k+1} \leftarrow v_k - \eta_k \cdot g_k$ ▷ update step
  $v_{k+1} \leftarrow v_{k+1} / \|v_{k+1}\|_2$ ▷ retraction to unit sphere
end for
return Final iterate v

3.2 Convergence Analysis

This section establishes convergence to first-order stationarity for Algorithm 1 under both loss functions. At any point where L is differentiable, a point $v \in \mathbb{S}^{d-1}$ is a first-order stationary point if $\operatorname{grad} L(v) = 0$ (Boumal et al., 2019). On the unit sphere, $\operatorname{grad} L(v) = (I - vv^\top)\nabla L(v)$. Specifically, we establish convergence under a diminishing learning rate for the Spearman loss and under a sufficiently small constant learning rate for the Fenchel–Young loss. We use the iteration notation of Algorithm 1 and introduce the following additional notation.

Notation Let $F := [f_1 \quad \cdots \quad f_n]^\top \in \mathbb{R}^{n\times d}$, so that $s(v) = Fv$. Let $s_{(1)}(v) \le \cdots \le s_{(n)}(v)$ denote the ordered scores of $s_1(v), \ldots, s_n(v)$, and write $s_{(\cdot)}(v) := \big(s_{(1)}(v), \ldots, s_{(n)}(v)\big)^\top$.

For a score vector $s \in \mathbb{R}^n$, some of the inequality constraints defining the permutahedron are satisfied with equality at the soft rank solution $r_n(s)$. We refer to the collection of these constraints as the active set. As s varies, the active set may change, inducing different regions in the score space. Let $\Sigma \subset \mathbb{R}^n$ denote the boundaries between these regions. A formal definition of Σ and the corresponding regions is provided in Section A.2.

Theorem 1 (Convergence of Algorithm 1 with Spearman loss)**.** *Let $\{v_k\}$ be generated by Algorithm 1 with L_n^{Sp} and learning rates satisfying*

$$\eta_k > 0,\ \eta_k \to 0 \text{ as } k \to \infty \text{ and } \sum_{k=1}^{\infty} \eta_k = \infty.$$

Let Ω be the set of accumulation points of $\{v_k\}$. Assume that $F\Omega \cap \Sigma = \emptyset$. Then, for all sufficiently large k,

$$L_n^{Sp}(v_{k+1}) \le L_n^{Sp}(v_k) - \frac{\eta_k}{4}\|g_k\|_2^2.$$

Moreover, $\|g_k\|_2 \to 0$ as $k \to \infty$.

The assumption $F\Omega \cap \Sigma = \emptyset$ excludes accumulation on the boundaries between active set regions.

For the Fenchel–Young loss, the smoothness of the objective leads to a simpler convergence result.

Theorem 2 (Convergence of Algorithm 1 with Fenchel–Young loss)**.** *Let $\{v_k\}$ be generated by Algorithm 1 with L_n^{FY} and constant learning rate $\eta_k \equiv \eta$ satisfying $0 < \eta \le 1/M$, where M is defined in Section A.2. Then,*

$$L_n^{FY}(v_{k+1}) \le L_n^{FY}(v_k) - \frac{\eta}{2}\|g_k\|_2^2.$$

Moreover, $\|g_k\|_2 \to 0$ as $k \to \infty$.

The Fenchel–Young loss differs from the Spearman loss in that it is differentiable everywhere with a Lipschitz continuous gradient, allowing convergence to be established without the boundary condition in Theorem 1. In both cases, $\|g_k\| \to 0$, establishing convergence to first-order stationarity.

The proofs of Theorems 1 and 2 are provided in Section A.2.

3.3 Consistency

We next study the statistical consistency of the rank axis estimators as the sample size increases. Unlike the convergence results in the previous subsection, the results in this section concern global minimizers of the empirical objectives. For either loss, let

$$\hat{v}_n \in \operatorname*{arg\,min}_{v\in\mathbb{S}^{d-1}} L_n(v). \tag{4}$$

Our analysis follows the standard M-estimation framework, based on uniform convergence to an appropriate population criterion and uniqueness of its minimizer.

We impose the following assumptions.

(A1) For any $v \in \mathbb{S}^{d-1}$, $v^\top f$ has a continuous distribution whose cumulative distribution function is strictly increasing on its support.

(A2) Let $F_{v^\top f}$ denote the cumulative distribution function of $v^\top f$. The quantile function $Q_v(u) := F^{-1}_{v^\top f}(u), u \in [0,1]$ is continuous.

(A3) For fixed $v \in \mathbb{S}^{d-1}$, as $n \to \infty$,

$$\max_{1 \le i \le n} \left| s_{(i)}(v) - Q_v\left(\frac{i}{n}\right) \right| \to 0 \quad \text{a.s.}$$

(A4) For every $v \in \mathbb{S}^{d-1}$, $v^\top f$ has density p_v satisfying $p_v(x) \le K$.

(A5) The soft-rank regularization parameter is $l_n = \gamma/n$, where $\gamma = 1/K$.

(A6) The observations satisfy $f_i \overset{\text{i.i.d.}}{\sim} P_f$ and $a_i = \langle v^*, f_i \rangle + \varepsilon_i$, where P_f is supported on a compact subset of $\mathbb{R}^d$, $v^* \in \mathbb{S}^{d-1}$, $\varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$, and $\{\varepsilon_i\}$ is independent of $\{f_i\}$.

Assumptions (A1)–(A3) provide regularity of the projected score distributions and connect empirical order statistics to their population quantiles. Assumption (A4) imposes a uniform upper bound on the projected score densities. Assumption (A5) specifies the $1/n$ scaling of the soft rank regularization parameter. As n increases, the gaps between adjacent ordered projection scores typically decrease, motivating a regularization parameter l_n of the same order so that the amount of smoothing does not increase relative to the rank spacing. Assumption (A6) specifies the data-generating model with Gaussian noise.

Under these assumptions, we first obtain almost sure consistency for the Spearman loss.

Theorem 3 (Consistency with Spearman loss)**.** *Under Assumptions (A1)–(A6), any*

$$\hat{v}_n \in \operatorname*{arg\,min}_{v \in \mathbb{S}^{d-1}} L_n^{Sp}(v)$$

satisfies $\hat{v}_n \to v^*$ almost surely.

For the Fenchel–Young loss, the same conclusion is established in the noise-free setting.

Theorem 4 (Consistency with Fenchel–Young loss)**.** *Under Assumptions (A1)–(A5) and (A6) with $\sigma = 0$, any*

$$\hat{v}_n \in \operatorname*{arg\,min}_{v \in \mathbb{S}^{d-1}} L_n^{FY}(v)$$

satisfies $\hat{v}_n \to v^*$ almost surely.

The proofs are provided in Section A.3, using the standard M-estimation result underlying Theorem 2.1 in Newey & McFadden (1994).

The restriction to the noise-free setting in Theorem 4 is not merely technical. The true axis need not minimize the population Fenchel–Young loss in the presence of Gaussian noise. In contrast, Theorem 3 establishes consistency of the Spearman loss under Gaussian noise. This difference motivates the use of the nonsmooth Spearman loss despite the smoothness of the Fenchel–Young loss. The following example illustrates that an incorrect axis can attain a lower population Fenchel–Young loss than the true axis.

Example 1 (The true axis need not minimize the Fenchel–Young loss under Gaussian noise)**.** *Let $f = (2X, Y) \in \mathbb{R}^2$, where $X, Y \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(0,1)$, let $\varepsilon \sim N(0,1)$, and let $v^* = (1,0)$ be the true axis. Define*

$$S_v := \langle v, f \rangle, \quad U_v := F_{S_v}(S_v), \quad T_* := F_{S_{v^*}+\varepsilon}(S_{v^*} + \varepsilon),$$

where F_Z denotes the cumulative distribution function of a random variable Z. For $v \in \mathbb{S}^1$, the population Fenchel–Young loss is

$$L^{FY}(v) = \mathbb{E}\left[\frac{2}{\gamma} S_v(U_v - T_*) - (U_v^2 - T_*^2)\right] = \frac{2}{\gamma}\left(\mathbb{E}[S_v U_v] - \mathbb{E}[S_v T_*]\right),$$

since U_v and T_* are uniformly distributed on $(0,1)$. Then, for $v' = (0,1)$,

$$L^{FY}(v^*) > L^{FY}(v').$$

Complete computational details and numerical demonstrations for this example are provided in Section A.4.

4 NUMERICAL STUDY

We design experiments to evaluate the effectiveness of the proposed rank axis learning framework in both synthetic and real-world settings. Synthetic data experiments examine whether the estimated axis recovers the ground-truth direction under varying sample sizes, embedding dimensions, and noise levels. In real datasets, where no ground-truth axis is available, we instead evaluate the rankability of the learned direction, i.e., how well its projected scores preserve the ordering of the target attribute.

4.1 SYNTHETIC DATA: GROUND-TRUTH AXIS RECOVERY

A ground-truth axis v^* is sampled uniformly from the unit sphere $\mathbb{S}^{d-1}$. We generate synthetic embeddings $f_i \sim N(0, I_d)$ and define the corresponding attribute as $a_i = v^{*\top} f_i + \varepsilon_i$ where $\varepsilon_i \sim N(0, \sigma^2)$. We vary the training sample size over $n \in \{50, 100, 200, 500, 1000\}$, the embedding dimension over $d \in \{32, 64, 128\}$, and the noise level over $\sigma \in \{0.0, 0.5\}$.

Since the true axis v^* is known, we evaluate axis recovery using the cosine similarity $\cos(\hat{v}, v^*)$ and the ranking performance using Spearman's rank correlation coefficient (SRCC). For $x, y \in \mathbb{R}^n$, we define the SRCC by $\rho(x, y) := \mathrm{Corr}\left(\mathrm{rank}(x), \mathrm{rank}(y)\right)$. We report the SRCC between the projected scores and attribute values on the test set.

We compare five methods: soft rank estimator with the Spearman loss, soft rank estimator with the Fenchel–Young loss, soft top-k estimators with Spearman and Fenchel–Young losses, and ListNet (Cao et al., 2007). We refer to the first four methods as SRank-Sp, SRank-FY, STopk-Sp, and STopk-FY, respectively. For STopk-Sp and STopk-FY, we use the soft top-k operator of Sander et al. (2023). It is based on the capped simplex generated by permutations of a vector with k ones and $n - k$ zeros, and we set $k = n/2$. Motivated by the learning rate conditions in our convergence results, SRank-Sp and STopk-Sp use a diminishing learning rate $\eta_j = \eta/\sqrt{j+1}$, and SRank-FY and STopk-FY employ a constant learning rate. ListNet is trained with a constant learning rate.

We repeat the experiment 100 times and plot the mean cosine similarity and SRCC with ± 1 standard deviation. The full experiment details are provided in Section A.1.1.

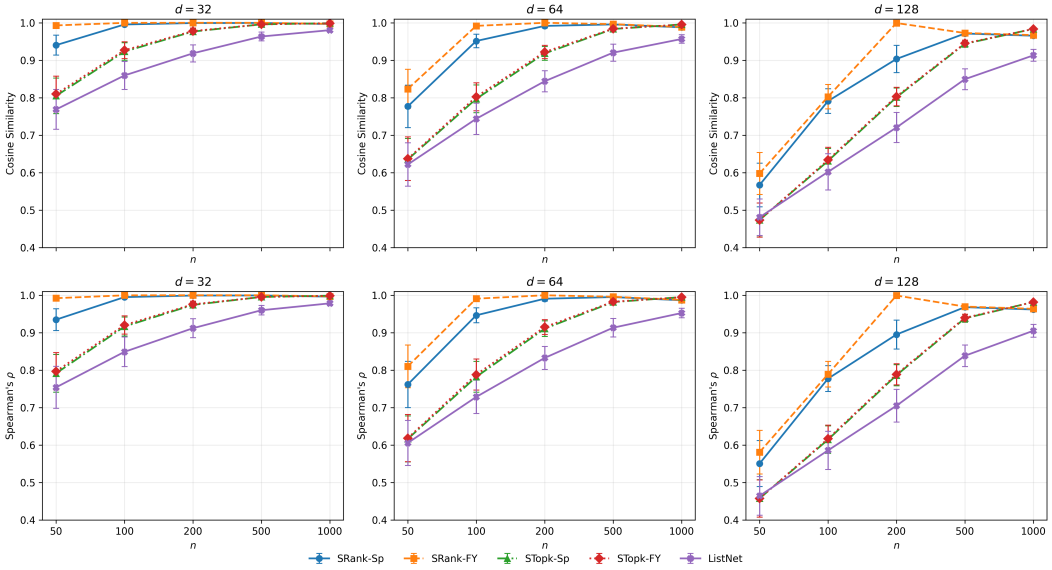


Figure 1: Cosine similarity and SRCC in the noise-free setting. Curves show the mean over 100 independent runs, with error bars indicating ± 1 standard deviation.

As shown in Figure 1, SRank-FY achieves the strongest overall performance, followed by SRank-Sp. The SRank methods generally outperform the STopk methods and ListNet. These results indicate that exploiting the full ranking structure is more effective for recovering the ground-truth axis than

the top-k or ListNet objectives in this synthetic setting. For all methods, cosine similarity and SRCC improve with increasing sample size and decreasing dimension. Adding noise reduces overall performance, while the relative advantage of the SRank methods remains clear. Here, we report the results for the noise-free setting. The noisy setting results are provided in Figure 2 of Section A.1.1.

4.2 Benchmark Datasets: Attribute-based Evaluation

We evaluate the proposed methods on image datasets. We use UTKFace (Zhang et al., 2017) and Adience (Eidinger et al., 2014) for age, UCF-QNRF (Idrees et al., 2018) and ShanghaiTech A/B (ST-A, B) (Zhang et al., 2016) for crowd counting, and KonIQ-10k (Hosu et al., 2020) for aesthetic quality. We extract pretrained visual embeddings using ResNet50 (He et al., 2016), ViT-B/32 (Dosovitskiy et al., 2021), ConvNeXtV2 (Woo et al., 2023), DINOv2 (ViT-B/14) (Oquab et al., 2024), CLIP ResNet50, CLIP ViT-B/32, and CLIP ConvNeXtV2 (Radford et al., 2021). Details of the datasets and embedding models are provided in Section A.1.2.

The main evaluation metric is Spearman's rank correlation coefficient (SRCC) between the projected scores and attribute values. We additionally evaluate top-k performance using precision, Jaccard similarity, AUC, and within-top-k Spearman correlation. Let T_k^{pred} and T_k^{gt} be the predicted and ground-truth top-k sets, respectively. We define $\text{Prec@}k := |T_k^{\text{pred}} \cap T_k^{\text{gt}}|/k$ and $\text{Jacc@}k := |T_k^{\text{pred}} \cap T_k^{\text{gt}}|/|T_k^{\text{pred}} \cup T_k^{\text{gt}}|$. AUC_k measures the probability that a randomly selected ground-truth top-k observation receives a higher score than a randomly selected non-top-k observation. $\text{Spearman@}k$ measures Spearman's rank correlation between the predicted scores and the attribute values within T_k^{gt} and is abbreviated as $\text{Spr@}k$ in the tables. These metrics are reported at $k = 5\%, 10\%$, and 20%.

We compare SRank-Sp, SRank-FY, STopk-Sp, and STopk-FY with linear regression (Sonthalia et al., 2025) and ListNet (Cao et al., 2007). The four rank and top-k based methods follow the same definitions as in Section 4.1.

4.3 Results

We first compare the overall ranking performance using SRCC, followed by a detailed evaluation of top-k performance. Table 1 reports the SRCC for each dataset, averaged over seven embedding models.

Table 1: Mean SRCC across seven embedding models for each benchmark dataset

Dataset	SRank -Sp	SRank -FY	STopk -Sp	STopk -FY	Linear Regression	ListNet
UTKFace	0.7770	0.7666	0.7691	0.7559	0.7638	0.6908
Adience	0.9092	0.9058	0.8476	0.8507	0.9094	0.8813
UCF_QNRF	0.8665	0.8706	0.8426	0.8308	0.6575	0.2931
ST-A	0.7653	0.7471	0.6799	0.7253	0.4369	0.4878
ST-B	0.8848	0.8785	0.8217	0.8181	0.8025	0.6129
KonIQ-10k	0.7996	0.7977	0.7814	0.7454	0.7828	0.6793

Overall, SRank-Sp achieves strong performance across the benchmark datasets, attaining the highest mean SRCC on 'UTKFace', 'ST-A', 'ST-B', 'KonIQ-10k'. SRank-FY performs best on 'UCF-QNRF', while linear regression achieves the highest SRCC on 'Adience' by a small margin. Compared with linear regression, SRank-Sp shows substantial improvements on 'UCF-QNRF', 'ST-A', and 'ST-B', and it outperforms ListNet on all six datasets.

We further examine whether the estimated axes accurately identify and order samples with the highest attribute values. Table 2 reports $\text{Prec@}k$, $\text{Jacc@}k$, AUC_k, and $\text{Spearman@}k$ for $k = 5\%, 10\%, 20\%$. The top-k results show a similar overall pattern. The SRank methods achieve the best performance on 11 of the 12 reported metrics, with SRank-Sp attaining the highest value on 6 metrics and SRank-FY on 5. SRank-FY performs well on the AUC metrics, indicating strong separation between samples inside and outside the ground-truth top-k set. In contrast, SRank-Sp achieves the best performance on most precision, Jaccard, and within-top-k Spearman metrics, sug-

Table 2: Average Top-k performance across all benchmark datasets and embedding models for $k \in \{5\%, 10\%, 20\%\}$

Metric	SRank Sp	SRank FY	STopk Sp	STopk FY	Linear Regression	ListNet
Prec@5%	0.4605	0.4525	0.4001	0.3642	0.3912	0.3312
Prec@10%	0.5946	0.5833	0.5224	0.4960	0.5055	0.4419
Prec@20%	0.6868	0.6896	0.6452	0.6347	0.6065	0.5284
Jacc@5%	0.3056	0.2988	0.2552	0.2289	0.2548	0.2055
Jacc@10%	0.4316	0.4207	0.3620	0.3356	0.3558	0.2924
Jacc@20%	0.5280	0.5308	0.4818	0.4692	0.4467	0.3659
$AUC_{5\%}$	0.9258	0.9268	0.8995	0.8951	0.8672	0.8471
$AUC_{10\%}$	0.9252	0.9268	0.9014	0.8974	0.8731	0.8357
$AUC_{20\%}$	0.9171	0.9176	0.8972	0.8940	0.8657	0.8129
Spr@5%	0.0827	0.0486	0.0002	0.0002	0.0773	0.0938
Spr@10%	0.2999	0.2730	0.2189	0.1777	0.2176	0.2011
Spr@20%	0.4764	0.4685	0.3895	0.3541	0.3769	0.3243

gesting that directly optimizing rank agreement is effective not only for recovering top-ranked samples but also for preserving their relative ordering. Notably, both SRank variants generally outperform STopk methods on these metrics despite using the full ranking structure during training. This suggests that global ordinal information can also be beneficial for performance near the top of the ranking. The only exception is Spr@5%, for which ListNet achieves the highest value. The relatively low Spr@5% values suggest that fine-grained ordering within the extreme top portion is more challenging than separating top-ranked samples from non-top-ranked samples, as also reflected by the substantially higher $AUC_{5\%}$ values.

In addition, we examine optimization convergence and runtime on the ShanghaiTech-A dataset using DINOv2 embeddings, with details provided in Appendix A.1.2. The training losses of the proposed methods decrease rapidly and stabilize within the first 300 epochs, indicating stable optimization well before the maximum training budget. In terms of runtime, the Fenchel–Young methods are slightly faster than their Spearman counterparts. In this setting, the proposed methods are substantially faster than linear regression, although slower than ListNet.

5 Conclusion

Our results show that rank axes can be effectively estimated by directly optimizing ordinal agreement rather than relying on the numerical scale of attribute labels. The proposed framework provides a practical approach to exploit ordinal structure in embedding spaces for representation analysis and attribute-aware retrieval. Our theoretical analysis establishes convergence to first-order stationarity and consistency under appropriate assumptions, while experiments demonstrate effective axis recovery and ranking performance across synthetic and real-world settings.

Several limitations remain. First, the behavior of the soft-rank operator depends on the regularization parameter relative to the gaps between adjacent projected scores. We mitigate this sensitivity by selecting the regularization parameter using validation data, but developing a principled selection rule with theoretical guarantees remains an important direction for future work. Second, our convergence results establish convergence to first-order stationarity rather than to a global minimizer. This is consistent with general nonconvex optimization, where global optimality typically requires additional assumptions and first-order methods are commonly analyzed in terms of convergence to stationary points (Jain & Kar, 2017; Boumal et al., 2019). In particular, although the Fenchel–Young loss is convex with respect to the score vector, the rank-axis problem optimizes the projection direction over the unit sphere and is therefore nonconvex. Consequently, standard global optimality guarantees for convex optimization do not directly apply. Future work may focus on more efficient optimization procedures, principled selection of the soft-rank regularization parameter, and conditions under which stronger global optimality guarantees can be established.

AI USE STATEMENT

In this work, we employed generative AI tools to refine the manuscript, assist with simple tasks related to experiment scripts, and perform minor checks of proofs. All AI-assisted content was carefully reviewed and validated by the authors to ensure its accuracy. The authors take full responsibility for the final content of this work, including text, claims, or artifacts produced with the assistance of generative AI.

REPRODUCIBILITY STATEMENT

The synthetic data distributions are specified in Section 4.1. Additional details on synthetic data generation, hyperparameters, and the training procedure are provided in Section A.1.1. The real-world datasets and their references are introduced in Section 4.2. Further details of real-world datasets and embedding models, preprocessing, hyperparameters, and the training procedure are provided in Section A.1.2.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations (Workshop Track)*, 2017. URL https://openreview.net/forum?id=ryF7rTqgl.

R. E. Barlow and H. D. Brunk. The isotonic regression problem and its dual. *Journal of the American Statistical Association*, 67(337):140–147, 1972. doi: 10.1080/01621459.1972.10481216.

Mathieu Blondel, André F.T. Martins, and Vlad Niculae. Learning with Fenchel-Young losses. *Journal of Machine Learning Research*, 21(35):1–69, 2020a. URL http://jmlr.org/papers/v21/19-021.html.

Mathieu Blondel, Olivier Teboul, Quentin Berthet, and Josip Djolonga. Fast differentiable sorting and ranking. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 950–959. PMLR, 2020b. URL https://proceedings.mlr.press/v119/blondel20a.html.

Nicolas Boumal, Pierre-Antoine Absil, and Coralia Cartis. Global rates of convergence for nonconvex optimization on manifolds. *IMA Journal of Numerical Analysis*, 39(1):1–33, 2019. doi: 10.1093/imanum/drx080.

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine Learning*, pp. 89–96, 2005. doi: 10.1145/1102351.1102363.

Wenzhi Cao, Vahid Mirjalili, and Sebastian Raschka. Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, 140:325–331, 2020. ISSN 0167-8655. doi: 10.1016/j.patrec.2020.11.008.

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th International Conference on Machine Learning*, pp. 129–136, 2007. doi: 10.1145/1273496.1273513.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

Koby Crammer and Yoram Singer. Pranking with ranking. In *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001. URL https://proceedings.neurips.cc/paper_files/paper/2001/file/5531a5834816222280f20d1ef9e95f69-Paper.pdf.

Marco Cuturi, Olivier Teboul, and Jean-Philippe Vert. Differentiable ranking and sorting using optimal transport. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/d8c24ca8f23c562a5600876ca2a550ce-Paper.pdf.

Jan Dhaene, Michel Denuit, Marc J. Goovaerts, Rob Kaas, and David Vyncke. The concept of comonotonicity in actuarial science and finance: Theory. *Insurance: Mathematics and Economics*, 31(1):3–33, 2002. doi: 10.1016/S0167-6687(02)00134-8.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Eran Eidinger, Roee Enbar, and Tal Hassner. Age and gender estimation of unfiltered faces. *IEEE Transactions on Information Forensics and Security*, 9(12):2170–2179, 2014. doi: 10.1109/TIFS.2014.2359646.

Martin Engilberge, Louis Chevallier, Patrick Pérez, and Matthieu Cord. SoDeep: A sorting deep net to learn ranking loss surrogates. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10784–10793, 2019. doi: 10.1109/CVPR.2019.01105.

Yoav Freund, Raj Iyer, Robert E. Schapire, and Yoram Singer. An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research*, 4:933–969, 2003. URL https://www.jmlr.org/papers/v4/freund03a.html.

Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2002–2011, 2018. doi: 10.1109/CVPR.2018.00214.

Aditya Grover, Eric Wang, Aaron Zweig, and Stefano Ermon. Stochastic optimization of sorting networks via continuous relaxations. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=H1eSS3CcKX.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Ralf Herbrich, Thore Graepel, and Klaus Obermayer. Large margin rank boundaries for ordinal regression. In *Advances in Large-Margin Classifiers*. The MIT Press, 2000. doi: 10.7551/mitpress/1113.003.0010.

Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29:4041–4056, 2020. doi: 10.1109/TIP.2020.2967829.

Haroon Idrees, Muhmmad Tayyab, Kishan Athrey, Dong Zhang, Somaya Al-Maadeed, Nasir Rajpoot, and Mubarak Shah. Composition loss for counting, density map estimation and localization in dense crowds. In *Proceedings of the European Conference on Computer Vision*, pp. 544–559. Springer, 2018. doi: 10.1007/978-3-030-01216-8_33.

Prateek Jain and Purushottam Kar. Non-convex optimization for machine learning. *Foundations and Trends in Machine Learning*, 10(3–4):142–336, 2017. doi: 10.1561/2200000058.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2668–2677. PMLR, 2018. URL https://proceedings.mlr.press/v80/kim18d.html.

Tie-Yan Liu. *Learning to Rank for Information Retrieval*. Springer, 2011. doi: 10.1007/978-3-642-14267-3.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=aajyHYjjsk.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf.

Hyeon-Woo Nam, Ye-Bin Moon, Lim Sohwi, Byung-Ki Kwon, and Tae-Hyun Oh. Zero-shot rankability: Revealing latent ordinal structure in multimodal large language models via language. In *International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=RsXgArgxC4.

Whitney K. Newey and Daniel McFadden. Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, volume 4, pp. 2111–2245. Elsevier, 1994. doi: 10.1016/S1573-4412(05)80005-4.

Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Ordinal regression with multiple output cnn for age estimation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4920–4928, 2016. doi: 10.1109/CVPR.2016.532.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

Devi Parikh and Kristen Grauman. Relative attributes. In *IEEE International Conference on Computer Vision*, pp. 503–510, 2011. doi: 10.1109/ICCV.2011.6126281.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 39643–39666. PMLR, 2024. URL https://proceedings.mlr.press/v235/park24c.html.

Alexander Postnikov. Permutohedra, associahedra, and beyond. *International Mathematics Research Notices*, 2009:1026–1106, 2009. doi: 10.1093/imrn/rnn153.

Sebastian Prillo and Julian Eisenschlos. SoftSort: A continuous relaxation for the argsort operator. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7793–7802. PMLR, 2020. URL https://proceedings.mlr.press/v119/prillo20a.html.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Michael Eli Sander, Joan Puigcerver, Josip Djolonga, Gabriel Peyré, and Mathieu Blondel. Fast, differentiable and sparse top-k: a convex analysis perspective. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29919–29936. PMLR, 2023. URL https://proceedings.mlr.press/v202/sander23a.html.

Krishna Kumar Singh and Yong Jae Lee. End-to-end localization and ranking for relative attributes. In *European Conference on Computer Vision*, pp. 753–769. Springer, 2016. doi: 10.1007/978-3-319-46466-4_45.

Ankit Sonthalia, Arnas Uselis, and Seong Joon Oh. On the rankability of visual embeddings. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 66169–66203. Curran Associates, Inc., 2025. doi: 10.52202/085713-2222. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/5f96a21345c138da929e99871fda138e-Paper-Conference.pdf.

Michael Taylor, John Guiver, Stephen Robertson, and Tom Minka. SoftRank: optimizing non-smooth rank metrics. In *International Conference on Web Search and Data Mining*, pp. 77–86, 2008. doi: 10.1145/1341531.1341544.

Rui Wang, Peipei Li, Huaibo Huang, Chunshui Cao, Ran He, and Zhaofeng He. Learning-to-rank meets language: Boosting language-driven ordering alignment for ordinal classification. In *Advances in Neural Information Processing Systems*, volume 36, pp. 76908–76922. Curran Associates, Inc., 2023. doi: 10.52202/075280-3361. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/f2a11632520f4b7473d7838f074a7d25-Paper-Conference.pdf.

Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. ConvNeXt V2: Co-designing and scaling convnets with masked autoencoders. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16133–16142, 2023. doi: 10.1109/CVPR52729.2023.01548.

Fen Xia, Tie-Yan Liu, Jue Wang, Wensheng Zhang, and Hang Li. Listwise approach to learning to rank - theory and algorithm. In *Proceedings of the 25th International Conference on Machine Learning*, pp. 1192–1199, 2008. doi: 10.1145/1390156.1390306. URL https://icml.cc/Conferences/2008/papers/167.pdf.

Wei-Hsiang Yu, Yen-Yu Lin, Ming-Hsuan Yang, and Yi-Hsuan Tsai. Ranking-aware adapter for text-driven image ordering with CLIP. In *International Conference on Learning Representations*, volume 2025, pp. 80614–80633, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/c86d3f43371270c70ed94e024852895d-Paper-Conference.pdf.

Kaiwen Zha, Peng Cao, Jeany Son, Yuzhe Yang, and Dina Katabi. Rank-N-Contrast: Learning continuous representations for regression. In *Advances in Neural Information Processing Systems*, volume 36, pp. 17882–17903. Curran Associates, Inc., 2023. doi: 10.52202/075280-0786. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/39e9c5913c970e3e49c2df629daff636-Paper-Conference.pdf.

Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 589–597, 2016. doi: 10.1109/CVPR.2016.70.

Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4352–4360, 2017. doi: 10.1109/CVPR.2017.463.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency, 2023. URL https://arxiv.org/abs/2310.01405.

A Appendix

A.1 Numerical study

A.1.1 Synthetic Data

This section provides experiment details and additional experiment results omitted from Section 4.1.

Experiment Details The validation and test sets, containing $1,000$ and $2,000$ observations, respectively, are generated from the same distribution as the training data.

The regularization parameter l and the learning rate η for the soft-rank or soft-top-k operator are selected using the validation set. We select η based on the validation loss, and then select l based on the validation SRCC. The candidate values are $l \in \{1, 0.5, 0.2, 0.1, 0.05, 0.02, 0.01, 0.005, 0.001\}$ and $\eta \in \{1, 0.1, 0.01, 0.001\}$. We use the implementation of the soft-top-k operator provided by Sander et al. (2023), available at `https://github.com/google-research/google-research/tree/master/sparse_soft_topk`. The original implementation is designed for soft top-k, whose feasible set is the capped simplex, which is the convex hull of permutations of a vector with k ones and $n - k$ zeros. To implement soft rank, we replace this vector with $(1, \ldots, n)$, whose permutations generate the permutahedron $\mathcal{P}(n)$ used in our formulation.

We train each model up to $3,000$ epochs. We evaluate the validation loss every 50 epochs, and training is stopped early if it fails to improve by at least 10^{-4} for 20 consecutive evaluations. We retain the model with the lowest validation loss.

Additional Results While Section 4.1 focuses on the noise-free setting, we additionally report results with $\sigma = 0.5$. Figure 2 shows the cosine similarity and SRCC under this noisy setting. As in the noise-free setting shown in Figure 1, SRank outperforms ListNet.

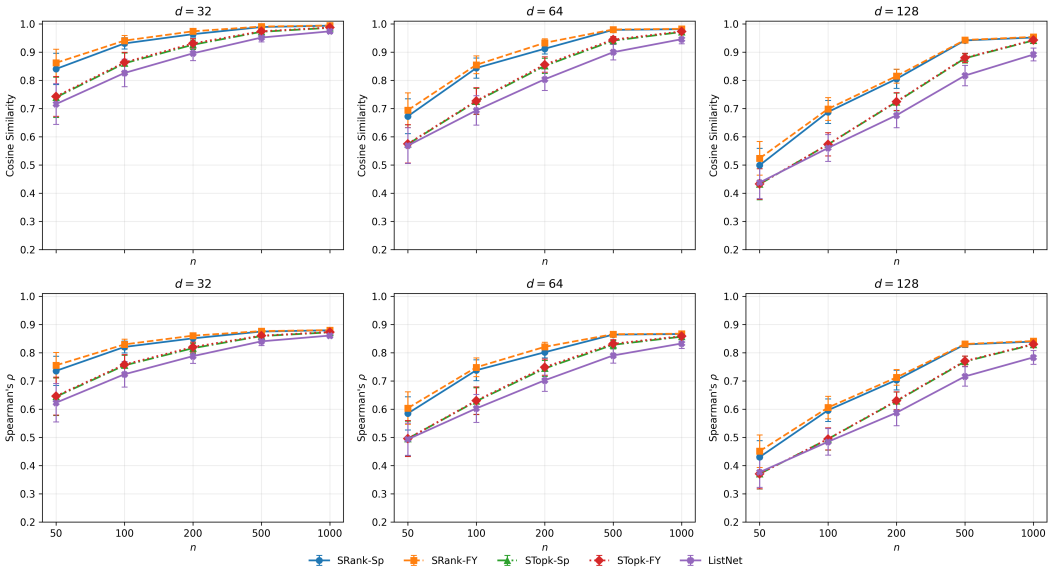


Figure 2: Cosine similarity and SRCC in the noisy setting. Curves show the mean over 100 independent runs, with error bars indicating ± 1 standard deviation.

A.1.2 Real Data

This section provides experimental details, optimization convergence and runtime analyzes, and additional results omitted from Section 4.3.

Experiment Details Table 3 summarizes the real-world datasets and ordinal attributes used in our experiments.

Following Sonthalia et al. (2025), we use the same set of seven vision encoders: ResNet50 (RN50), ViT-B/32 (ViTB32), ConvNeXtV2 (CNX), DINOv2 ViT-B/14 (DINO-B14), CLIP ResNet50 (CLIP-RN50), CLIP ViT-B/32 (CLIP-ViTB32), and CLIP ConvNeXtV2 (CLIP-CNX). We extract the image embeddings using the code released by Sonthalia et al. (2025), available at `https://github.com/aktsonthalia/rankable-vision-embeddings`. Before optimization, we center the embeddings using the mean of the training set and scale them by the average norm of the centered training embeddings. We apply the same transformation to the validation and test sets.

Table 3: Datasets and ordinal attributes used for rank-axis evaluation. Each dataset provides a scalar attribute a_i defining the ranking task.

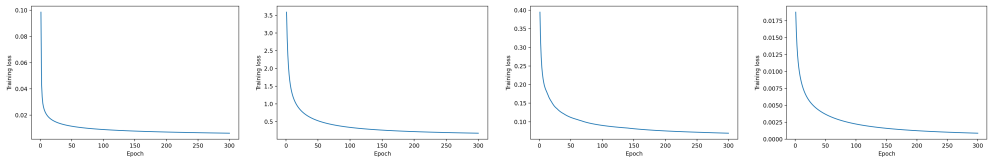
Dataset	Task	Attribute (a_i)
UTKFace	Age ranking	chronological age
Adience	Age group ranking	ordinal age group
UCF-QNRF	Crowd ranking	crowd count
ST-A	Crowd ranking	crowd count
ST-B	Crowd ranking	crowd count
KonIQ-10k	Image quality ranking	MOS

This standardizes the embedding mean overall scale while preserving the ordering induced by any fixed projection direction.

We use the same implementations of the soft-rank and soft-top-k operators and the same training procedure as in the synthetic experiments. The candidate values for the regularization parameter l and the learning rate η differ from those used in the synthetic experiments:

$$l \in \{1\} \cup \{c \times 10^{-k} : c \in \{1, 2, 5\}, k = 1 \ldots, 6\}, \quad \eta \in \{1, 0.5, 0.1, 0.05, 0.01, 0.005, 0.0001\}.$$

Optimization Convergence and Runtime We plot the training loss curves of the four proposed methods on ShanghaiTech-A dataset using DINOv2 embeddings in Figure 3. We show only the first 300 epochs because the losses change only marginally thereafter, although all methods were trained for up to 3000 epochs. For all four methods, the training loss decreases rapidly and then stabilizes within the first 300 epochs, indicating stable optimization well before the maximum training epochs.



(a) SRank-Sp (b) SRank-FY (c) STopk-Sp (d) STopk-FY

Figure 3: Training loss curves on ShanghaiTech-A dataset with DINOv2 embeddings

We also measure the runtime on the same ShanghaiTech-A/DINOv2 setting using an NVIDIA TITAN V GPU. For each method, runtime is measured using the selected hyperparameter setting, excluding the hyperparameter search. The runtimes are 5.306s for SRank-Sp, 4.073s for SRank-FY, 5.360s for STopk-Sp, 4.131s for STopk-FY, 16.032s for linear regression, and 2.295s for ListNet. The four proposed methods have comparable runtimes. The methods using the Fenchel–Young loss are slightly faster than their Spearman counterparts. In this setting, they are substantially faster than linear regression, although slower than ListNet.

Additional Results In Section 4.3, we report the SRCC averaged over the seven embedding models for each dataset. Here, we provide the detailed results for each combination of dataset and embedding model. Table 4 reports the results for SRank-Sp, Table 5 for SRank-FY, Table 6 for STopk-Sp, and Table 7 for STopk-FY, respectively. The highest SRCC for each dataset and embedding model is highlighted. Overall, the detailed results are consistent with the averaged results in Section 4.3. The SRank methods generally achieve higher SRCC across embedding models.

Table 4: SRCC for SRank-Sp

Dataset	RN50	ViTB32	CNX	DINO -B14	CLIP -RN50	CLIP -ViTB32	CLIP -CNX
UTKFace	0.6681	0.7648	0.7911	0.7745	0.8015	0.8120	0.8269
Adience	0.8355	0.8892	0.9035	0.9154	0.9308	0.9444	0.9455
UCF_QNRF	0.8565	0.8727	0.8686	0.8427	0.8927	0.8670	0.8655
ST-A	0.7813	0.8247	0.7771	0.6529	0.7293	0.7996	0.7920
ST-B	0.9016	0.8883	0.8742	0.8509	0.8929	0.8878	0.8977
KonIQ-10k	0.7571	0.7502	0.7805	0.7483	0.8480	0.8370	0.8761

Table 5: SRCC for SRank-FY

Dataset	RN50	ViTB32	CNX	DINO -B14	CLIP -RN50	CLIP -ViTB32	CLIP -CNX
UTKFace	0.6552	0.7182	0.7803	0.7684	0.8030	0.8141	0.8266
Adience	0.8270	0.8859	0.8994	0.9077	0.9315	0.9430	0.9458
UCF_QNRF	0.8726	0.8726	0.8687	0.8560	0.8735	0.8806	0.8699
ST-A	0.8356	0.8320	0.7894	0.7386	0.7737	0.6466	0.6137
ST-B	0.8896	0.8547	0.8760	0.8715	0.9029	0.8806	0.8739
KonIQ-10k	0.7542	0.7500	0.7782	0.7504	0.8481	0.8264	0.8763

Table 6: SRCC for STopk-Sp

Dataset	RN50	ViTB32	CNX	DINO -B14	CLIP -RN50	CLIP -ViTB32	CLIP -CNX
UTKFace	0.6594	0.7599	0.7841	0.7705	0.7920	0.8031	0.8148
Adience	0.8188	0.8504	0.8560	0.8528	0.8536	0.8466	0.8548
UCF_QNRF	0.8534	0.8445	0.8398	0.7974	0.8511	0.8585	0.8536
ST-A	0.7247	0.7606	0.7225	0.4970	0.7204	0.5730	0.7613
ST-B	0.8709	0.7986	0.7838	0.8200	0.8216	0.7990	0.8579
KonIQ-10k	0.7354	0.7429	0.7597	0.7183	0.8306	0.8190	0.8643

Table 7: SRCC for STopk-FY

Dataset	RN50	ViTB32	CNX	DINO -B14	CLIP -RN50	CLIP -ViTB32	CLIP -CNX
UTKFace	0.6283	0.7503	0.7723	0.7502	0.7852	0.7950	0.8096
Adience	0.8172	0.8552	0.8547	0.8298	0.8608	0.8682	0.8691
UCF_QNRF	0.8429	0.8461	0.8278	0.8232	0.8291	0.8214	0.8250
ST-A	0.8050	0.7965	0.7425	0.6404	0.7374	0.6910	0.6640
ST-B	0.8811	0.8450	0.7769	0.7825	0.8370	0.7588	0.8454
KonIQ-10k	0.7001	0.6855	0.7291	0.6799	0.7951	0.7823	0.8460

Table 9: SRCC for ListNet

Dataset	RN50	ViTB32	CNX	DINO -B14	CLIP -RN50	CLIP -ViTB32	CLIP -CNX
UTKFace	0.5698	0.6900	0.6906	0.6661	0.7199	0.7253	0.7740
Adience	0.8240	0.8653	0.8723	0.8681	0.8971	0.9145	0.9275
UCF_QNRF	0.5390	0.2535	0.0819	0.1273	0.3888	0.3516	0.3094
ST-A	0.6195	0.6916	0.5570	0.4974	0.3224	0.3663	0.3603
ST-B	0.6092	0.5697	0.5788	0.6176	0.6092	0.6949	0.6109
KonIQ-10k	0.6415	0.6140	0.6722	0.6779	0.7030	0.6774	0.7693

Table 8: SRCC for Linear Regression

Dataset	RN50	ViTB32	CNX	DINO -B14	CLIP -RN50	CLIP -ViTB32	CLIP -CNX
UTKFace	0.6531	0.7515	0.7731	0.7550	0.7974	0.8014	0.8150
Adience	0.8337	0.8961	0.9017	0.9094	0.9331	0.9442	0.9473
UCF_QNRF	0.7161	0.7783	0.7937	0.6300	0.5491	0.5449	0.5900
ST-A	0.0849	0.5863	0.4124	0.6772	0.4301	0.4195	0.4475
ST-B	0.7995	0.8715	0.8719	0.8497	0.6241	0.8086	0.7924
KonIQ-10k	0.7330	0.7482	0.7378	0.7464	0.8249	0.8256	0.8634

A.2 Convergence Proofs

Proof of Theorem 1

Proof. (Step 1: Permutahedron representation) Write $[n] := \{1, \dots, n\}$ and define

$$\mathcal{T}_n := \{S \subset [n] : 1 \le |S| \le n-1\}.$$

For $S \in \mathcal{T}_n$, let

$$A_S := \mathbf{1}_S^\top, \quad b_S := \sum_{j=n-|S|+1}^{n} j = \frac{|S|(2n-|S|+1)}{2}.$$

Then we can write the permutahedron $\mathcal{P}(n)$ as

$$\mathcal{P}(n) = \{y \in \mathbb{R}^n : A_S y \le b_S \text{ for all } S \in \mathcal{T}_n, \mathbf{1}^\top y = c_n\}, \quad \text{where } c_n := \frac{n(n+1)}{2}.$$

Consider a chain of subsets of $[n]$

$$\mathcal{T} = (S_1, \dots, S_r), \quad \emptyset \subsetneq S_1 \subsetneq \cdots \subsetneq S_r \subsetneq [n].$$

Here, $r = 0$ is allowed, in which case $\mathcal{T} = \emptyset$. Let $\mathcal{C}$ be the collection of all such nested chains. For $\mathcal{T} = (S_1, \dots, S_r) \in \mathcal{C}$, define

$$A_\mathcal{T} := \begin{bmatrix} A_{S_1}^\top & \cdots & A_{S_r}^\top \end{bmatrix}^\top, \quad b_\mathcal{T} := \begin{bmatrix} b_{S_1} & \cdots & b_{S_r} \end{bmatrix}^\top,$$

with the convention that $A_\mathcal{T} \in \mathbb{R}^{0\times n}$ and $b_\mathcal{T} \in \mathbb{R}^0$ when $r = 0$. Define

$$C_\mathcal{T} := \begin{bmatrix} A_\mathcal{T}^\top & \mathbf{1} \end{bmatrix}^\top, \quad d_\mathcal{T} := \begin{bmatrix} b_\mathcal{T}^\top & c_n \end{bmatrix}^\top.$$

By Postnikov (2009), the chains in $\mathcal{C}$ index the faces of $\mathcal{P}(n)$, whose active equalities are $C_\mathcal{T} y = d_\mathcal{T}$.

(Step 2: $C_\mathcal{T} C_\mathcal{T}^\top$ is invertible) Since the sets in $\mathcal{T}$ are strictly nested, $\mathbf{1}_{S_1}^\top, \dots, \mathbf{1}_{S_r}^\top, \mathbf{1}^\top$ are linearly independent. Hence $C_\mathcal{T}$ has full row rank, so $C_\mathcal{T} C_\mathcal{T}^\top$ is invertible.

(Step 3: KKT conditions) For fixed $s \in \mathbb{R}^n$, the problem defining the soft rank $r_n(s)$ is convex with affine constraints. Thus, the KKT conditions are necessary and sufficient for optimality. The Lagrangian of the optimization problem is

$$\mathcal{L}(y, \nu, \tau) = \frac{l_n}{2}\|y\|_2^2 - s^\top y + \sum_{S \in \mathcal{T}_n} \nu_S (A_S y - b_S) + \tau(\mathbf{1}^\top y - c_n),$$

where $\nu = (\nu_S)_{S \in \mathcal{T}_n} \ge 0$ and $\tau \in \mathbb{R}$ are the Lagrange multipliers. The KKT conditions are

$$\begin{aligned} &\nabla_y \mathcal{L}(y, \nu, \tau) = l_n y - s + \sum_{S \in \mathcal{T}_n} \nu_S A_S^\top + \tau \mathbf{1} = 0, \\ &A_S y - b_S \le 0, \quad \nu_S \ge 0, \quad \nu_S (A_S y - b_S) = 0 \quad \text{for every } S \in \mathcal{T}_n, \\ &\mathbf{1}^\top y - c_n = 0. \end{aligned}$$

Fix $\mathcal{T} = (S_1, \ldots, S_r) \in \mathcal{C}$. We consider the linear system obtained by taking the constraints indexed by $S_1, \ldots, S_r$ as active and setting the remaining inequality multipliers to zero. Then

$$l_n y - s + A_{\mathcal{T}}^\top \nu_{\mathcal{T}} + \tau \mathbf{1} = 0, \quad A_{\mathcal{T}} y = b_{\mathcal{T}}, \quad \mathbf{1}^\top y = c_n,$$

where $\nu_{\mathcal{T}} := (\nu_{S_1}, \ldots, \nu_{S_r})^\top$. This system can be written as

$$\begin{bmatrix} l_n I & C_{\mathcal{T}}^\top \\ C_{\mathcal{T}} & O \end{bmatrix} \begin{bmatrix} y \\ \lambda_{\mathcal{T}} \end{bmatrix} = \begin{bmatrix} s \\ d_{\mathcal{T}} \end{bmatrix}, \quad \lambda_{\mathcal{T}} := \begin{bmatrix} \nu_{\mathcal{T}} \\ \tau \end{bmatrix}.$$

Since $y = (s - C_{\mathcal{T}}^\top \lambda_{\mathcal{T}})/l_n$ and $C_{\mathcal{T}} y = d_{\mathcal{T}}$, we have

$$d_{\mathcal{T}} = C_{\mathcal{T}} y = \frac{1}{l_n}(C_{\mathcal{T}} s - C_{\mathcal{T}} C_{\mathcal{T}}^\top \lambda_{\mathcal{T}}).$$

Since $C_{\mathcal{T}} C_{\mathcal{T}}^\top$ is invertible,

$$\lambda_{\mathcal{T}} = (C_{\mathcal{T}} C_{\mathcal{T}}^\top)^{-1} C_{\mathcal{T}} s - l_n (C_{\mathcal{T}} C_{\mathcal{T}}^\top)^{-1} d_{\mathcal{T}}.$$

Since s is arbitrary, we write $\lambda_{\mathcal{T}}(s)$ and $\nu_{\mathcal{T}}(s)$ to make their dependence on s explicit. Define $y_{\mathcal{T}}(s) := (s - C_{\mathcal{T}}^\top \lambda_{\mathcal{T}}(s))/l_n$. Then

$$y_{\mathcal{T}}(s) = \frac{1}{l_n} Q_{\mathcal{T}} s + c_{\mathcal{T}}, \tag{5}$$

where $Q_{\mathcal{T}} := I - C_{\mathcal{T}}^\top (C_{\mathcal{T}} C_{\mathcal{T}}^\top)^{-1} C_{\mathcal{T}}$ and $c_{\mathcal{T}} := C_{\mathcal{T}}^\top (C_{\mathcal{T}} C_{\mathcal{T}}^\top)^{-1} d_{\mathcal{T}}$.

(Step 4: Active set regions $\mathcal{R}_{\mathcal{T}}$) The inactive slacks $b_S - A_S y_{\mathcal{T}}(s)$ for $S \in \mathcal{T}_n \setminus \{S_1, \ldots, S_r\}$ and the components of $\nu_{\mathcal{T}}(s)$ are affine in s. The active set region $\mathcal{R}_{\mathcal{T}}$ associated with $\mathcal{T} = (S_1, \ldots, S_r)$ is defined by

$$\mathcal{R}_{\mathcal{T}} := \bigcap_{S \in \mathcal{T}_n \setminus \{S_1, \ldots, S_r\}} \{s : b_S - A_S y_{\mathcal{T}}(s) > 0\} \cap \bigcap_{k=1}^{r} \{s : [\nu_{\mathcal{T}}(s)]_k > 0\}.$$

For $r = 0$, the second intersection is understood as $\mathbb{R}^n$. Since $\mathcal{R}_{\mathcal{T}}$ is an intersection of finitely many open affine half spaces, it is open. For $s \in \mathcal{R}_{\mathcal{T}}$, define the multipliers for all inequality constraints by

$$\nu_S(s) := \begin{cases} [\nu_{\mathcal{T}}(s)]_k, & S = S_k \text{ for some } k = 1, \ldots, r \\ 0, & S \notin \{S_1, \ldots, S_r\}. \end{cases}$$

For $S = S_k$ with $k = 1, \ldots, r$, $A_S y_{\mathcal{T}}(s) = b_S$ and $\nu_S(s) > 0$. For $S \notin \{S_1, \ldots, S_r\}$, $A_S y_{\mathcal{T}}(s) < b_S$ and $\nu_S(s) = 0$. Therefore, for every $S \in \mathcal{T}_n$,

$$A_S y_{\mathcal{T}}(s) - b_S \le 0, \quad \nu_S(s) \ge 0, \quad \nu_S(s)(A_S y_{\mathcal{T}}(s) - b_S) = 0.$$

Together with the linear system in Step 3, these show that all KKT conditions are satisfied. Hence

$$r_n(s) = y_{\mathcal{T}}(s) \quad \text{for every } s \in \mathcal{R}_{\mathcal{T}}.$$

Finally, the regions $\mathcal{R}_{\mathcal{T}}$ are pairwise disjoint. Let $\mathcal{T} \neq \mathcal{T}'$, where

$$\mathcal{T} = (S_1, \ldots, S_r), \quad \mathcal{T}' = (S'_1, \ldots, S'_{r'}).$$

Then there exists a set that belongs to one chain but not the other. Without loss of generality, suppose that

$$S \in \{S_1, \ldots, S_r\} \quad \text{and} \quad S \notin \{S'_1, \ldots, S'_{r'}\}.$$

If $s \in \mathcal{R}_{\mathcal{T}}$, then $A_S r_n(s) = b_S$. On the other hand, if $s \in \mathcal{R}_{\mathcal{T}'}$, then $A_S r_n(s) < b_S$. Therefore,

$$\mathcal{R}_{\mathcal{T}} \cap \mathcal{R}_{\mathcal{T}'} = \emptyset.$$

(Step 5: Boundedness of the Riemannian gradient) Let $J_{\mathcal{T}}$ be the Jacobian matrix of $y_{\mathcal{T}}(s)$. Then

$$J_{\mathcal{T}} = \frac{\partial y_{\mathcal{T}}}{\partial s} = \frac{1}{l_n} Q_{\mathcal{T}}. \tag{6}$$

Since $Q_{\mathcal{T}}$ is the orthogonal projector onto $\ker(C_{\mathcal{T}})$, we have

$$\|J_{\mathcal{T}}\|_2 \leq \frac{1}{l_n}. \tag{7}$$

For fixed $\mathcal{T}$, let

$$L_{\mathcal{T}}(v) = \frac{1}{n^3} \left\| y_{\mathcal{T}}(Fv) - t^{(n)} \right\|_2^2$$

define the Spearman loss on $\{v : s(v) \in \mathcal{R}_{\mathcal{T}}\}$. Then the gradient of $L_{\mathcal{T}}(v)$ is

$$\nabla L_{\mathcal{T}}(v) = \frac{2}{n^3} F^\top J_{\mathcal{T}}^\top (y_{\mathcal{T}}(Fv) - t^{(n)})$$

and the Riemannian gradient of $L_{\mathcal{T}}(v)$ is

$$g_{\mathcal{T}}(v) := \operatorname{grad} L_{\mathcal{T}}(v) = (I - vv^\top)\nabla L_{\mathcal{T}}(v).$$

Since $\|J_{\mathcal{T}}\|_2 \leq 1/l_n$ (equation 7) and $y_{\mathcal{T}}(Fv), t^{(n)} \in \mathcal{P}(n)$, there exists G, independent of $\mathcal{T}$ and v, such that $\|g_{\mathcal{T}}(v)\|_2 \leq G$ whenever $Fv \in \mathcal{R}_{\mathcal{T}}$. Thus, $\|g_k\|_2 \leq G$ whenever $Fv_k \in \bigcup_{\mathcal{T}\in\mathcal{C}} \mathcal{R}_{\mathcal{T}}$. At a boundary point, if the loss is differentiable, its Riemannian gradient is the limit of the gradients from any incident differentiable region. Otherwise, g_k is given by such a limiting gradient from one of the incident differentiable regions, as specified in Algorithm 1. Consequently, the same bound holds.

(Step 6: Update path) If $g_k \neq 0$, define $u_k := g_k/\|g_k\|_2$. Since $g_k \in T_{v_k}\mathbb{S}^{d-1}$, we have $\langle u_k, v_k\rangle = 0$. For $g_k \neq 0$, the update rule for $\{v_k\}$ can be written as

$$v_{k+1} = \cos\theta_k v_k - \sin\theta_k u_k, \quad \theta_k := \arctan(\eta_k \|g_k\|_2).$$

If $g_k = 0$, set $\theta_k = 0$. Then we have

$$0 \leq \theta_k \leq \eta_k \|g_k\|_2 \leq G\eta_k.$$

Since $\eta_k \to 0$, we have $\theta_k \to 0$.

(Step 7: The path is contained in one region for sufficiently large k) The sphere $\mathbb{S}^{d-1}$ is compact, so the set of accumulation points Ω is nonempty and compact. Since each $\mathcal{R}_{\mathcal{T}}$ is open by Step 4, $\Sigma := \mathbb{R}^n \setminus \bigcup_{\mathcal{T}\in\mathcal{C}} \mathcal{R}_{\mathcal{T}}$ is closed. Since $F\Omega$ is compact, Σ is closed, and $F\Omega \cap \Sigma = \emptyset$,

$$\delta := \operatorname{dist}(F\Omega, \Sigma) > 0.$$

To relate the iterates to the accumulation set Ω, we first show that $\operatorname{dist}(v_k, \Omega) \to 0$. Suppose not. Then there exist $\varepsilon > 0$ and a subsequence $\{v_{k_j}\}$ with $\operatorname{dist}(v_{k_j}, \Omega) \geq \varepsilon$. By compactness, a further subsequence converges to some $\bar{v} \in \Omega$, a contradiction. Thus

$$\operatorname{dist}(Fv_k, F\Omega) \leq \|F\|_2 \operatorname{dist}(v_k, \Omega) \to 0.$$

Therefore,

$$\operatorname{dist}(Fv_k, \Sigma) > \frac{3\delta}{4} \text{ for sufficiently large } k.$$

If $g_k \neq 0$, define the path $v_k(t) := (\cos t)v_k - (\sin t)u_k$ for $0 \leq t \leq \theta_k$. For $g_k = 0$, use the constant path. Define $s_k(t) := Fv_k(t)$. For every $t \in [0, \theta_k]$,

$$\|s_k(t) - s_k(0)\|_2 \leq \|F\|_2 \|v_k(t) - v_k\|_2 = 2\|F\|_2 \sin\frac{t}{2} \leq \|F\|_2 \theta_k.$$

Since $\theta_k \to 0$ by Step 6,

$$\|F\|_2 \theta_k < \frac{\delta}{4} \text{ for sufficiently large } k.$$

Thus, for sufficiently large k, the following holds for every $t \in [0, \theta_k]$.

$$\operatorname{dist}(s_k(t), \Sigma) \geq \operatorname{dist}(s_k(0), \Sigma) - \|s_k(t) - s_k(0)\|_2 > \frac{3\delta}{4} - \frac{\delta}{4} = \frac{\delta}{2}.$$

Hence, the update path does not intersect Σ for sufficiently large k. Therefore, for sufficiently large k, the connected path $s_k([0, \theta_k])$ is contained in $\mathbb{R}^n \setminus \Sigma = \bigcup_{\mathcal{T}\in\mathcal{C}} \mathcal{R}_{\mathcal{T}}$. Since the regions $\mathcal{R}_{\mathcal{T}}$ are pairwise disjoint by Step 4, this path is contained in one region $\mathcal{R}_{\mathcal{T}_k}$. The endpoint of the path

$s_k(\theta_k) = Fv_{k+1}$ is in $\mathcal{R}_{\mathcal{T}_k}$ and $\mathcal{R}_{\mathcal{T}_{k+1}}$, so $\mathcal{T}_k = \mathcal{T}_{k+1}$ since the regions are pairwise disjoint. Therefore, there exist k_1 and $\mathcal{T}^* \in \mathcal{C}$ such that

$$Fv_k(t) \in \mathcal{R}_{\mathcal{T}^*} \text{ for all } k \geq k_1 \text{ and } 0 \leq t \leq \theta_k.$$

(Step 8: Analysis in the fixed region $\mathcal{R}_{\mathcal{T}^*}$) For fixed $k \geq k_1$, define $\phi_k(t) := L_n^{\text{Sp}}(v_k(t))$ for $t \in [0, \theta_k]$. Since $s_k(t) \in \mathcal{R}_{\mathcal{T}^*}$ by Step 7,

$$L_n^{\text{Sp}}(v_k(t)) = L_{\mathcal{T}^*}(v_k(t)) = \frac{1}{n^3} \left\| y_{\mathcal{T}^*}(s_k(t)) - t^{(n)} \right\|_2^2.$$

By equation 6 and $s_k''(t) = -s_k(t)$,

$$\frac{d}{dt} y_{\mathcal{T}^*}(s_k(t)) = J_{\mathcal{T}^*} s_k'(t), \quad \frac{d^2}{dt^2} y_{\mathcal{T}^*}(s_k(t)) = J_{\mathcal{T}^*} s_k''(t) = -J_{\mathcal{T}^*} s_k(t),$$

and

$$\begin{aligned}
\phi_k'(t) &= \frac{2}{n^3} \left\langle \frac{d}{dt} y_{\mathcal{T}^*}(s_k(t)), y_{\mathcal{T}^*}(s_k(t)) - t^{(n)} \right\rangle, \\
\phi_k''(t) &= \frac{2}{n^3} \left(\left\langle \frac{d}{dt} y_{\mathcal{T}^*}(s_k(t)), \frac{d}{dt} y_{\mathcal{T}^*}(s_k(t)) \right\rangle + \left\langle \frac{d^2}{dt^2} y_{\mathcal{T}^*}(s_k(t)), y_{\mathcal{T}^*}(s_k(t)) - t^{(n)} \right\rangle \right) \\
&= \frac{2}{n^3} \left(\|J_{\mathcal{T}^*} s_k'(t)\|_2^2 - \left(y_{\mathcal{T}^*}(s_k(t)) - t^{(n)} \right)^\top J_{\mathcal{T}^*} s_k(t) \right).
\end{aligned}$$

Since $\|J_{\mathcal{T}^*}\|_2 \leq 1/l_n$ by equation 7, $\|s_k(t)\|_2$ and $\|s_k'(t)\|_2$ are bounded, and $y_{\mathcal{T}^*}(s_k(t)), t^{(n)} \in \mathcal{P}(n)$, there exists $H > 0$ independent of k and t such that $\phi_k''(t) \leq H$.

(Step 9: Loss descent) Since $\eta_k \to 0$ and $\|g_k\|_2$ is bounded, choose $k_2 \geq k_1$ such that $H\eta_k \leq 1$ and $\eta_k \|g_k\|_2 \leq 1$ for all $k \geq k_2$. Fix $k \geq k_2$ and suppose that $g_k \neq 0$. The chain rule on $\mathbb{S}^{d-1}$ gives $\phi_k'(0) = \langle g_k, v_k'(0) \rangle$. Since $v_k'(0) = -u_k = -g_k / \|g_k\|_2$,

$$\phi_k'(0) = \left\langle g_k, -\frac{g_k}{\|g_k\|_2} \right\rangle = -\|g_k\|_2.$$

The integral form of Taylor's theorem gives

$$L_n^{\text{Sp}}(v_{k+1}) - L_n^{\text{Sp}}(v_k) = \phi_k(\theta_k) - \phi_k(0) \leq -\theta_k \|g_k\|_2 + \frac{H}{2} \theta_k^2.$$

Since $\theta_k = \arctan(\eta_k \|g_k\|_2) \leq \eta_k \|g_k\|_2$,

$$\frac{H}{2} \theta_k^2 \leq \frac{1}{2} \theta_k \|g_k\|_2.$$

Thus

$$L_n^{\text{Sp}}(v_{k+1}) \leq L_n^{\text{Sp}}(v_k) - \frac{1}{2} \theta_k \|g_k\|_2.$$

Since $0 \leq \eta_k \|g_k\|_2 \leq 1$ and $\arctan x \geq x/2$ for $0 \leq x \leq 1$, $\theta_k \geq \eta_k \|g_k\|_2 / 2$ and

$$L_n^{\text{Sp}}(v_{k+1}) \leq L_n^{\text{Sp}}(v_k) - \frac{\eta_k}{4} \|g_k\|_2^2, \tag{8}$$

for $k \geq k_2$. The same inequality holds when $g_k = 0$. Since $L_n^{\text{Sp}}(v) \geq 0$, summing equation 8 gives

$$\sum_{k=k_2}^{\infty} \eta_k \|g_k\|_2^2 < \infty.$$

(Step 10: Convergence of $\|g_k\|_2$) We extend $L_{\mathcal{T}^*}$ to the ambient space $\mathbb{R}^d$. The extended loss function is

$$\tilde{L}_n^{\text{Sp}}(v) := \frac{1}{n^3} \|J_{\mathcal{T}^*} F v + c_{\mathcal{T}^*} - t^{(n)}\|_2^2.$$

Define

$$\tilde{g}(v) := (I - vv^\top) \nabla \tilde{L}_n^{\text{Sp}}(v), \quad v \in \mathbb{R}^d.$$

Since $\tilde{g}$ is continuously differentiable, it is Lipschitz on $\mathbb{S}^{d-1}$. Hence there exists $L_g > 0$ such that

$$\|\tilde{g}(v) - \tilde{g}(w)\|_2 \leq L_g \|v - w\|_2 \text{ for all } v, w \in \mathbb{S}^{d-1}.$$

For $k \geq k_1$, $g_k = \tilde{g}(v_k)$ and

$$|\|g_{k+1}\|_2 - \|g_k\|_2| \leq \|g_{k+1} - g_k\|_2 \leq L_g \|v_{k+1} - v_k\|_2 \leq L_g \theta_k \leq L_g \eta_k \|g_k\|_2.$$

Let $a_k := \|g_k\|_2$. Then $|a_{k+1} - a_k| \leq L_g \eta_k a_k$.

Suppose $a_k \nrightarrow 0$. Since $a_k \geq 0$, there exists $\varepsilon' > 0$ such that $a_k \geq \varepsilon'$ infinitely often. Since $\sum_{k=k_2}^{\infty} \eta_k a_k^2 < \infty$ and $\sum \eta_k = \infty$, $a_k \leq \varepsilon'/2$ for infinitely many k. Choose disjoint intervals $[p_j, q_j]$ such that

$$a_{p_j} \leq \varepsilon'/2, \quad a_{q_j} \geq \varepsilon', \quad \varepsilon'/2 < a_k < \varepsilon' \text{ for } k \in (p_j, q_j).$$

Hence

$$\frac{\varepsilon'}{2} \leq a_{q_j} - a_{p_j} \leq \sum_{k=p_j}^{q_j - 1} |a_{k+1} - a_k| \leq L_g \sum_{k=p_j}^{q_j - 1} \eta_k a_k \leq L_g \varepsilon' \sum_{k=p_j}^{q_j - 1} \eta_k.$$

Therefore,

$$\sum_{k=p_j}^{q_j - 1} \eta_k \geq \frac{1}{2L_g}.$$

Since $p_j \to \infty$ and $\eta_k \to 0$, $\eta_{p_j} \leq 1/(6L_g)$ for sufficiently large j. Hence, for sufficiently large j,

$$\sum_{k=p_j+1}^{q_j - 1} \eta_k a_k^2 \geq \frac{\varepsilon'^2}{4} \sum_{k=p_j+1}^{q_j - 1} \eta_k \geq \frac{\varepsilon'^2}{12L_g}.$$

Since the intervals are disjoint, summing over j yields

$$\sum_{k=k_2}^{\infty} \eta_k a_k^2 = \infty,$$

a contradiction. Therefore, $\|g_k\|_2 = a_k \to 0$.

□

Proof of Theorem 2

Proof. By Proposition 2 of Blondel et al. (2020a), the quadratic regularization used in the soft rank and the chain rule imply that $L_n^{\text{FY}}(v)$ has Lipschitz continuous gradient. Since $\mathbb{S}^{d-1}$ is a compact submanifold of $\mathbb{R}^d$, Lemma 2.7 of Boumal et al. (2019) implies that Assumption 2.6 of Boumal et al. (2019) holds globally for some $L_g > 0$. Since $L_n^{\text{FY}}(v)$ is continuous on $\mathbb{S}^{d-1}$, it attains its minimum $L^* := \min_{v \in \mathbb{S}^{d-1}} L_n^{\text{FY}}(v)$. Set $M := L_g$ and fix $0 < \eta \leq 1/M$. Since $1/\eta \geq M = L_g$, Assumption 2.6 also holds with constant $1/\eta$. From the proof of Theorem 2.8 in Boumal et al. (2019),

$$L_n^{\text{FY}}(v_{k+1}) \leq L_n^{\text{FY}}(v_k) - \frac{\eta}{2} \|g_k\|_2^2.$$

Summing from $k = 0$ to k_1 gives

$$\frac{\eta}{2} \sum_{k=0}^{k_1} \|g_k\|_2^2 \leq L_n^{\text{FY}}(v_0) - L_n^{\text{FY}}(v_{k_1+1}) \leq L_n^{\text{FY}}(v_0) - L^*.$$

Letting $k_1 \to \infty$, we obtain

$$\sum_{k=0}^{\infty} \|g_k\|_2^2 < \infty,$$

and hence $\|g_k\|_2 \to 0$.

□

A.3 CONSISTENCY PROOFS

We first introduce several auxiliary results used in the proofs of both consistency theorems.

Notation For the consistency analysis, we make the dependence on the sample size explicit. For each n and $v \in \mathbb{S}^{d-1}$, let

$$s_{(1)}^{(n)}(v) \le s_{(2)}^{(n)}(v) \le \cdots \le s_{(n)}^{(n)}(v)$$

denote the sorted scores of $s_1(v), \ldots, s_n(v)$, and write $s_{(\cdot)}^{(n)}(v) := \left(s_{(1)}^{(n)}(v), \ldots, s_{(n)}^{(n)}(v)\right)^\top$. Define

$$y^{(n)}(v) := r_n\left(s_{(\cdot)}^{(n)}(v)\right).$$

Let $q^{(n)}(v) := (Q_v(1/n), Q_v(2/n), \ldots, Q_v(1))^\top$ and $\xi^{(n)} := (1, \ldots, n)^\top$.

Let $f \sim P_f$ and define $S_v := \langle v, f \rangle$. We write F_{S_v} for its cumulative distribution function and define $U_v := F_{S_v}(S_v)$. Let $\varepsilon \sim N(0, \sigma^2)$ be independent of f, and define $A := S_{v^*} + \varepsilon$. Let F_A be the cumulative distribution function of A, and define $T_* := F_A(A)$.

Define

$$\Delta_n(v) := \max_{1 \le k \le n} \left| s_{(k)}^{(n)}(v) - Q_v\left(\frac{k}{n}\right) \right|.$$

Under Assumption (A3), $\Delta_n(v) \to 0$ a.s.

Preliminary Technical Lemmas

Lemma 1 (Population quantiles and hard ranks)**.** *Under Assumptions (A1), (A2), (A4), and (A5), for every $n \ge 2$ and $v \in \mathbb{S}^{d-1}$,*

$$r_n\left(q^{(n)}(v)\right) = \xi^{(n)}.$$

Proof. By Assumptions (A1) and (A2), $F_{S_v}\left(q_i^{(n)}(v)\right) = i/n$ for $i = 1, \ldots, n$. Hence, by Assumption (A4), for $i = 1, \ldots, n-1$,

$$\frac{1}{n} = F_{S_v}\left(q_{i+1}^{(n)}(v)\right) - F_{S_v}\left(q_i^{(n)}(v)\right) \le K\left(q_{i+1}^{(n)}(v) - q_i^{(n)}(v)\right).$$

By Assumption (A5), $q_{i+1}^{(n)}(v) - q_i^{(n)}(v) \ge 1/(Kn) = l_n$, and Lemma 3 in Blondel et al. (2020b) implies $r_n\left(q^{(n)}(v)\right) = \xi^{(n)}$. □

Lemma 2. *Under Assumptions (A1), (A2) and (A4), for every $v \in \mathbb{S}^{d-1} \setminus \{\pm v^*\}$,*

$$P(U_v = U_{v^*}) < 1.$$

Proof. Suppose for contradiction that $U_v = U_{v^*}$ a.s. for some $v \in \mathbb{S}^{d-1} \setminus \{\pm v^*\}$. By Assumption (A1), F_{S_v} is strictly increasing on its support, so

$$S_v = Q_v(U_v) = Q_v(U_{v^*}) = Q_v(F_{S_{v^*}}(S_{v^*})) \quad \text{a.s.}$$

Define $g(x) := Q_v(F_{S_{v^*}}(x))$. Then $S_v = g(S_{v^*})$ a.s. Since $F_{S_{v^*}}$ and Q_v are nondecreasing, g is nondecreasing and hence differentiable almost everywhere. Since $\int_{\mathbb{R}} p_{v^*}(x)dx = 1$, the set $\{x : p_{v^*}(x) > 0\}$ has positive Lebesgue measure. By the Lebesgue differentiation theorem and the fact that g is differentiable almost everywhere, we may choose $x_0 \in \mathbb{R}$ such that

$$p_{v^*}(x_0) > 0, \; g'(x_0) = c$$

for some $c \in \mathbb{R}$, and

$$\lim_{r \to 0} \frac{1}{2r} \int_{x_0 - r}^{x_0 + r} p_{v^*}(t)dt = p_{v^*}(x_0).$$

Define

$$\delta_v := \|v - cv^*\|_2, \quad \xi := \frac{v - cv^*}{\delta_v}.$$

Since $v \notin \{\pm v^*\}$, we have $\delta_v > 0$, and hence $\xi \in \mathbb{S}^{d-1}$. Then

$$S_\xi = \langle \xi, f \rangle = \frac{S_v - cS_{v^*}}{\delta_v} = \frac{g(S_{v^*}) - cS_{v^*}}{\delta_v} \quad \text{a.s.}$$

Define $h(x) := g(x) - cx$. Then $h'(x_0) = 0$ and $h(S_{v^*}) = \delta_v S_\xi$ a.s. Since $h'(x_0) = 0$, for every $\alpha > 0$, there exists $R > 0$ such that

$$|x - x_0| < R \Longrightarrow |h(x) - h(x_0)| \le \alpha |x - x_0|.$$

Since $h(S_{v^*}) = \delta_v S_\xi$, for every $0 < r < R$,

$$|S_{v^*} - x_0| < r \Longrightarrow \left| S_\xi - \frac{h(x_0)}{\delta_v} \right| \le \frac{\alpha r}{\delta_v}.$$

Thus

$$P(|S_{v^*} - x_0| < r) \le P\left(\left| S_\xi - \frac{h(x_0)}{\delta_v} \right| \le \frac{\alpha r}{\delta_v} \right).$$

By the choice of x_0, there exists $R' > 0$ such that for every $0 < r < R'$,

$$\frac{1}{2r} \int_{x_0 - r}^{x_0 + r} p_{v^*}(t) dt \ge \frac{1}{2} p_{v^*}(x_0).$$

Therefore, for $0 < r < R'$,

$$P(|S_{v^*} - x_0| < r) = \int_{x_0 - r}^{x_0 + r} p_{v^*}(t) dt \ge p_{v^*}(x_0) r.$$

By Assumption (A4), $p_\xi \le K$, and hence

$$P\left(\left| S_\xi - \frac{h(x_0)}{\delta_v} \right| \le \frac{\alpha r}{\delta_v} \right) \le K \cdot \frac{2\alpha r}{\delta_v}.$$

Therefore, for $0 < r < \min(R, R')$, $p_{v^*}(x_0) r \le 2K\alpha r / \delta_v$, which gives $p_{v^*}(x_0) \le 2K\alpha / \delta_v$. Since $\alpha > 0$ is arbitrary, letting $\alpha \to 0^+$ gives $p_{v^*}(x_0) \le 0$, contradicting $p_{v^*}(x_0) > 0$. This contradiction proves that $P(U_v = U_{v^*}) < 1$. □

Lemma 3. *Let U_1 and U_2 have $\mathrm{Unif}(0,1)$ marginals, and let $g : [0,1] \to \mathbb{R}$ be a continuous and strictly increasing function. Then*

$$\mathbb{E}[U_1 g(U_2)] \le \mathbb{E}[U g(U)], \quad U \sim \mathrm{Unif}(0,1),$$

with equality if and only if $U_1 = U_2$ a.s.

Proof. Let $X_1 = U_1$ and $X_2 = g(U_2)$. Then their quantile functions are $F^{-1}_{X_1}(u) = u$ and $F^{-1}_{X_2}(u) = g(u)$ for $u \in (0,1)$. Hence, by Theorem 2 in Dhaene et al. (2002), the comonotonic counterpart is distributed as $(U, g(U))$ for $U \sim \mathrm{Unif}(0,1)$. By Theorem 8 in Dhaene et al. (2002),

$$\mathrm{Corr}(U_1, g(U_2)) \le \mathrm{Corr}(U, g(U))$$

and equality holds if and only if $(U_1, g(U_2))$ is comonotonic. Since U, U_1 and U_2 have the same distribution,

$$\mathbb{E}[U_1 g(U_2)] \le \mathbb{E}[U g(U)]$$

and equality holds if and only if $(U_1, g(U_2))$ is comonotonic.

Suppose that $(U_1, g(U_2))$ is comonotonic. Then by Theorem 2 in Dhaene et al. (2002),

$$(U_1, g(U_2)) \overset{d}{=} (F^{-1}_{U_1}(V), F^{-1}_{g(U_2)}(V)) \text{ for some } V \sim \mathrm{Unif}(0,1).$$

Since $F^{-1}_{U_1}(u) = u$ and $F^{-1}_{g(U_2)}(u) = g(u)$, $(U_1, g(U_2)) \overset{d}{=} (V, g(V))$. Applying the measurable map $(x, y) \mapsto g(x) - y$ gives $g(U_1) - g(U_2) \overset{d}{=} 0$, and thus $g(U_1) = g(U_2)$ a.s. Since g is strictly increasing, $U_1 = U_2$ a.s.

Conversely, if $U_1 = U_2$ a.s., then $\mathbb{E}[U_1 g(U_2)] = \mathbb{E}[U g(U)]$. □

Proof of Theorem 3 Let $L^{\mathrm{Sp}}(v) := \mathbb{E}\left[(U_v - T_*)^2\right]$ be the population Spearman loss. Throughout this proof, we write L and L_n for L^{Sp} and L_n^{Sp}, respectively, since no other loss is considered.

Lemma 4 (Pointwise convergence)**.** *Under Assumptions (A1)–(A6), for each fixed $v \in \mathbb{S}^{d-1}$,*

$$L_n^{Sp}(v) \to L^{Sp}(v) \quad a.s.$$

Proof. (Step 1: Soft rank approximation) Completing the square of the objective function of equation 1 gives

$$r_n(s(v)) = \arg\min_{y \in \mathcal{P}(n)} \left\| y - \frac{s(v)}{l_n} \right\|_2^2 = \mathrm{Proj}_{\mathcal{P}(n)}\left(\frac{s(v)}{l_n}\right).$$

Since the Euclidean projection onto the closed convex set $\mathcal{P}(n)$ is non-expansive, r_n is $1/l_n$-Lipschitz. By Lemma 1,

$$\begin{aligned}\left\| r_n\left(s_{(\cdot)}^{(n)}(v)\right) - \xi^{(n)} \right\|_2 &= \left\| \mathrm{Proj}_{\mathcal{P}(n)}\left(\frac{s_{(\cdot)}^{(n)}(v)}{l_n}\right) - \mathrm{Proj}_{\mathcal{P}(n)}\left(\frac{q^{(n)}(v)}{l_n}\right) \right\|_2 \\ &\le \frac{1}{l_n}\left\| s_{(\cdot)}^{(n)}(v) - q^{(n)}(v) \right\|_2\end{aligned}$$

Since $l_n = \gamma/n$ by Assumption (A5),

$$\begin{aligned}\frac{1}{n}\sum_{k=1}^{n}\left|\frac{y_k^{(n)}(v)}{n} - \frac{k}{n}\right| &\le \frac{1}{n^{3/2}}\left\| r_n\left(s_{(\cdot)}^{(n)}(v)\right) - \xi^{(n)} \right\|_2 \\ &\le \frac{1}{\sqrt{n}\gamma}\left\| s_{(\cdot)}^{(n)}(v) - q^{(n)}(v) \right\|_2 \\ &\le \frac{1}{\gamma}\Delta_n(v)\end{aligned}$$

Since S_v has a continuous distribution, the scores $s_1(v), \ldots, s_n(v)$ are distinct almost surely. Let π_n denote the corresponding ordering permutation, so that

$$s_{\pi_n(1)}(v) < \cdots < s_{\pi_n(n)}(v) \quad \text{a.s.}$$

For $i = 1, \ldots, n$, define $U_{v,i} := F_{S_v}(s_i(v))$. Since $s_{\pi_n(k)}(v) = s_{(k)}^{(n)}(v)$,

$$U_{v,\pi_n(k)} = F_{S_v}\left(s_{(k)}^{(n)}(v)\right).$$

By Assumptions (A1), (A2), and (A4),

$$\max_{1\le k\le n}\left|U_{v,\pi_n(k)} - \frac{k}{n}\right| = \max_{1\le k\le n}\left|F_{S_v}\left(s_{(k)}^{(n)}(v)\right) - F_{S_v}\left(Q_v\left(\frac{k}{n}\right)\right)\right| \le K\Delta_n(v).$$

Therefore,

$$\begin{aligned}\frac{1}{n}\sum_{k=1}^{n}\left|\frac{1}{n}y_k^{(n)}(v) - U_{v,\pi_n(k)}\right| &\le \frac{1}{n}\sum_{k=1}^{n}\left|\frac{1}{n}y_k^{(n)}(v) - \frac{k}{n}\right| + \max_{1\le k\le n}\left|\frac{k}{n} - U_{v,\pi_n(k)}\right| \\ &\le \left(\frac{1}{\gamma} + K\right)\Delta_n(v) = 2K\Delta_n(v) \to 0 \quad \text{a.s.}\end{aligned}$$

(Step 2: Approximation of the hard rank by the distribution function) Since A has a continuous distribution, there are no ties among $a_1, \ldots, a_n$ almost surely. Let

$$\hat{F}_n(x) := \frac{1}{n}\sum_{j=1}^{n}\mathbf{1}_{\{a_j \le x\}}$$

be the empirical distribution function. Then almost surely,

$$\frac{t_i^{(n)}}{n} = \hat{F}_n(a_i), \quad i = 1, \ldots, n.$$

Define $T_i := F_A(a_i)$. By the Glivenko–Cantelli theorem,

$$\max_{1\le i\le n}\left|\frac{t_i^{(n)}}{n}-T_i\right| = \max_{1\le i\le n}\left|\hat{F}_n(a_i)-F_A(a_i)\right| \le \sup_{x\in\mathbb{R}}\left|\hat{F}_n(x)-F_A(x)\right| \to 0 \quad \text{a.s.}$$

(Step 3: Approximation of the empirical loss) Define

$$\tilde{L}_n(v) := \frac{1}{n}\sum_{k=1}^{n}\left(U_{v,\pi_n(k)}-T_{\pi_n(k)}\right)^2.$$

By permutation invariance,

$$L_n(v) = \frac{1}{n^3}\left\|r_n(s(v))-t^{(n)}\right\|_2^2 = \frac{1}{n}\sum_{k=1}^{n}\left(\frac{1}{n}y_k^{(n)}(v)-\frac{1}{n}t^{(n)}_{\pi_n(k)}\right)^2.$$

Since $0 \le y_k^{(n)}(v)/n, t^{(n)}_{\pi_n(k)}/n, U_{v,\pi_n(k)}, T_{\pi_n(k)} \le 1$,

$$\begin{aligned}\left|L_n(v)-\tilde{L}_n(v)\right| &\le \frac{2}{n}\sum_{k=1}^{n}\left|\left(\frac{1}{n}y_k^{(n)}(v)-\frac{1}{n}t^{(n)}_{\pi_n(k)}\right)-\left(U_{v,\pi_n(k)}-T_{\pi_n(k)}\right)\right| \\ &\le \frac{2}{n}\sum_{k=1}^{n}\left|\frac{1}{n}y_k^{(n)}(v)-U_{v,\pi_n(k)}\right| + \frac{2}{n}\sum_{k=1}^{n}\left|\frac{1}{n}t^{(n)}_{\pi_n(k)}-T_{\pi_n(k)}\right| \\ &\le \frac{2}{n}\sum_{k=1}^{n}\left|\frac{1}{n}y_k^{(n)}(v)-U_{v,\pi_n(k)}\right| + 2\max_{1\le i\le n}\left|\frac{1}{n}t_i^{(n)}-T_i\right|.\end{aligned}$$

By Steps 1 and 2, $|L_n(v)-\tilde{L}_n(v)| \to 0$ a.s.

(Step 4: Convergence to the population loss) Reindexing the sum gives

$$\tilde{L}_n(v) = \frac{1}{n}\sum_{i=1}^{n}\left(U_{v,i}-T_i\right)^2.$$

The pairs $(U_{v,i}, T_i)$ are i.i.d. with the same joint distribution as

$$(U_v, T_*) = \left(F_{S_v}(S_v), F_A(S_{v^*}+\varepsilon)\right).$$

Since the summands are bounded, the strong law of large numbers yields

$$\tilde{L}_n(v) \to \mathbb{E}\left[(U_v-T_*)^2\right] = L(v) \quad \text{a.s.}$$

Combining this with Step 3, for each fixed $v \in \mathbb{S}^{d-1}$,

$$L_n(v) \to L(v) \quad \text{a.s.}$$

□

Lemma 5 (Uniform convergence)**.** *Under Assumptions (A1)–(A6),*

$$L_n^{Sp}(v) \to L^{Sp}(v) \quad \textit{uniformly on } \mathbb{S}^{d-1} \textit{ a.s.}$$

Proof. By Assumption (A6), there exist a constant $C > 0$ and an event Ω_C with $P(\Omega_C) = 1$ such that $\|f_i\|_2 \le C$ for all i on Ω_C. Then, on Ω_C,

$$\|F\|_2 \le \left(\sum_{i=1}^{n}\|f_i\|_2^2\right)^{1/2} \le C\sqrt{n}.$$

Let $c_n(v) := r_n(s(v))/n$ and $d_n := t^{(n)}/n$. Then

$$L_n(v) = \frac{1}{n}\left\|c_n(v)-d_n\right\|_2^2$$

and for any $v, w \in \mathbb{S}^{d-1}$,

$$|L_n(v) - L_n(w)| = \frac{1}{n} \left| \|c_n(v) - d_n\|_2^2 - \|c_n(w) - d_n\|_2^2 \right| \\ \leq \frac{1}{n} \|c_n(v) - c_n(w)\|_2 \left(\|c_n(v) - d_n\|_2 + \|c_n(w) - d_n\|_2 \right).$$

Since every component of $c_n(v)$ and d_n lies in $[0, 1]$,

$$\|c_n(v) - d_n\|_2 \leq \sqrt{n}, \quad \|c_n(w) - d_n\|_2 \leq \sqrt{n}.$$

As shown in Step 1 of Lemma 4, r_n is $1/l_n$-Lipschitz. On Ω_C,

$$\|r_n(s(v)) - r_n(s(w))\|_2 \leq \frac{1}{l_n} \|s(v) - s(w)\|_2 \leq \frac{\|F\|_2}{l_n} \|v - w\|_2,$$

and

$$\|c_n(v) - c_n(w)\|_2 = \frac{1}{n} \|r_n(s(v)) - r_n(s(w))\|_2 \leq \frac{C}{\sqrt{n} l_n} \|v - w\|_2.$$

Using $l_n = \gamma/n$,

$$|L_n(v) - L_n(w)| \leq \frac{1}{n} \times \frac{C}{\sqrt{n} l_n} \|v - w\|_2 \times 2\sqrt{n} = \frac{2C}{\gamma} \|v - w\|_2.$$

Thus, every L_n is $2C/\gamma$-Lipschitz on Ω_C. For fixed $v, w \in \mathbb{S}^{d-1}$, Lemma 4 implies that $L_n(v) \to L(v)$ and $L_n(w) \to L(w)$ simultaneously on a probability-one subset of Ω_C. On this event,

$$|L(v) - L(w)| = \lim_{n \to \infty} |L_n(v) - L_n(w)| \leq \frac{2C}{\gamma} \|v - w\|_2.$$

Since both sides are deterministic and v, w are arbitrary, L is $2C/\gamma$-Lipschitz. Moreover, $0 \leq L_n(v) \leq 1$ for all n and v, so $\{L_n\}$ is uniformly bounded.

Since $\mathbb{S}^{d-1}$ is compact, there exists a countable dense subset D. Since D is countable, we may choose an event $\Omega \subseteq \Omega_C$ of probability one such that $L_n(v) \to L(v)$ for every $v \in D$ on Ω. Fix $\omega \in \Omega$ and suppress the dependence on ω below. Let $\{L_{n_j}\}$ be an arbitrary subsequence. By the Arzelà–Ascoli theorem, it has a further subsequence that converges uniformly on $\mathbb{S}^{d-1}$ to some $\tilde{L}$. Since $L_n(v) \to L(v)$ for every $v \in D$, we have $\tilde{L}(v) = L(v)$ for every $v \in D$. Since $\tilde{L}$ and L are continuous and D is dense, $\tilde{L} = L$ on $\mathbb{S}^{d-1}$. Since every subsequence has a further subsequence that converges uniformly on $\mathbb{S}^{d-1}$ to L, the whole sequence converges uniformly to L. Hence,

$$\sup_{v \in \mathbb{S}^{d-1}} |L_n(v) - L(v)| \to 0.$$

Since $\omega \in \Omega$ is arbitrary and $P(\Omega) = 1$, the above convergence holds a.s.

□

Lemma 6 (Unique minimizer)**.** *Under Assumptions (A1)–(A6), $L^{Sp}(v)$ is uniquely minimized at v^*.*

Proof. Assume that $\sigma > 0$. Since S_v and $A = S_{v^*} + \varepsilon$ have continuous distributions, $U_v = F_{S_v}(S_v)$ and $T_* = F_A(A)$ are uniformly distributed on $(0, 1)$ by the probability integral transform. Then

$$L(v) = \mathbb{E}[(U_v - T_*)^2] = \mathbb{E}[U_v^2] + \mathbb{E}[T_*^2] - 2\mathbb{E}[U_v T_*] = \frac{2}{3} - 2\mathbb{E}[U_v T_*],$$

which implies that minimizing $L(v)$ is equivalent to maximizing $\mathbb{E}[U_v T_*]$. Since U_v is measurable with respect to f,

$$\mathbb{E}[U_v T_*] = \mathbb{E}[\mathbb{E}[U_v T_* | f]] = \mathbb{E}[U_v \mathbb{E}[T_* | f]] = \mathbb{E}[U_v \mathbb{E}[F_A(S_{v^*} + \varepsilon) | f]].$$

Define $g(s) := \mathbb{E}_\varepsilon[F_A(s + \varepsilon)]$ for $s \in \mathbb{R}$. Since ε is independent of f and S_{v^*} is measurable with respect to f, it follows that

$$\mathbb{E}[F_A(S_{v^*} + \varepsilon) | f] = g(S_{v^*}),$$

so that

$$\mathbb{E}[U_v T_*] = \mathbb{E}[U_v g(S_{v^*})].$$

Since $F_{S_{v^*}}$ is continuous and strictly increasing on its support, Q_{v^*} is its inverse on $(0,1)$. Hence,

$$S_{v^*} = Q_{v^*}(F_{S_{v^*}}(S_{v^*})) = Q_{v^*}(U_{v^*}) \quad \text{a.s.}$$

Define $h(u) := g(Q_{v^*}(u)) = \mathbb{E}[F_A(Q_{v^*}(u) + \varepsilon)]$. Since F_A is continuous and bounded, the dominated convergence theorem implies that g is continuous. Hence, $h = g \circ Q_{v^*}$ is continuous. Moreover,

$$g(S_{v^*}) = g(Q_{v^*}(U_{v^*})) = h(U_{v^*}) \quad \text{a.s.}$$

Therefore,

$$\mathbb{E}[U_v T_*] = \mathbb{E}[U_v h(U_{v^*})].$$

Now, we will show that h is strictly increasing. Since $F_{S_{v^*}}$ is continuous and strictly increasing on its support, Q_{v^*} is strictly increasing on $(0,1)$. Since $\sigma > 0$, A has a strictly positive density on $\mathbb{R}$, and hence F_A is strictly increasing on $\mathbb{R}$. Thus, for every $0 < u_1 < u_2 < 1$ and $c \in \mathbb{R}$,

$$F_A(Q_{v^*}(u_1) + c) < F_A(Q_{v^*}(u_2) + c).$$

Taking expectation with respect to ε, we have $h(u_1) < h(u_2)$. Since h is continuous, h is strictly increasing on $[0,1]$. By Lemma 3,

$$\mathbb{E}[U_v h(U_{v^*})] \le \int_0^1 uh(u)du$$

with equality if and only if $U_v = U_{v^*}$ almost surely. By Lemma 2, $P(U_v = U_{v^*}) < 1$ for every $v \in \mathbb{S}^{d-1} \setminus \{\pm v^*\}$, implying that $U_v = U_{v^*}$ cannot hold almost surely for any such v.

Since $S_{-v^*} = -S_{v^*}$ and S_{v^*} has a continuous distribution,

$$U_{-v^*} = F_{-S_{v^*}}(-S_{v^*}) = 1 - F_{S_{v^*}}(S_{v^*}) = 1 - U_{v^*} \quad \text{a.s.}$$

Hence, since $U_{v^*} \sim \mathrm{Unif}(0,1)$,

$$P(U_{-v^*} = U_{v^*}) = P\left(U_{v^*} = \frac{1}{2}\right) = 0.$$

Thus, $U_v = U_{v^*}$ a.s. cannot hold when $v = -v^*$. For $v = v^*$, we have $U_v = U_{v^*}$ a.s., so equality is attained. Therefore, v^* is the unique maximizer of $\mathbb{E}[U_v T_*] = \mathbb{E}[U_v h(U_{v^*})]$, which implies that v^* is the unique minimizer of L.

For $\sigma = 0$, we have $T_* = U_{v^*}$. Therefore, $L(v) = \mathbb{E}\left[(U_v - U_{v^*})^2\right]$. Clearly, $L(v^*) = 0$. For every $v \in \mathbb{S}^{d-1} \setminus \{\pm v^*\}$, Lemma 2 gives $P(U_v = U_{v^*}) < 1$, which implies $L(v) > 0$. As shown above, $\mathbb{P}(U_{-v^*} = U_{v^*}) = 0$, and hence $L(-v^*) > 0$. Therefore, $L(v)$ is uniquely minimized at v^*. □

Combining Lemmas 4, 5, and 6 with the strong consistency result for M-estimators (see Theorem 2.1 and the discussion immediately following it in Newey & McFadden (1994)), we conclude

$$\hat{v}_n^{\mathrm{Sp}} \to v^* \quad \text{a.s.}$$

Proof of Theorem 4 We rescale the empirical Fenchel–Young loss in equation 3 as

$$\bar{L}_n^{\mathrm{FY}}(v) := \frac{2}{\gamma n} L_n^{\mathrm{FY}}(v) = \frac{2}{\gamma n^2} \sup_{y \in \mathcal{P}(n)} \left(\langle s(v), y \rangle - \frac{l_n}{2}\|y\|_2^2 + \frac{l_n}{2}\left\|t^{(n)}\right\|_2^2 - \left\langle s(v), t^{(n)} \right\rangle \right).$$

Since the factor $2/(\gamma n)$ is positive and independent of v, this rescaling does not affect the minimizer. Hence, we analyze the rescaled empirical objective without loss of generality. We define the corresponding population Fenchel–Young loss by

$$L^{\mathrm{FY}}(v) := \mathbb{E}\left[\frac{2}{\gamma} S_v(U_v - T_*) - (U_v^2 - T_*^2)\right].$$

Throughout the proofs of Lemmas 7 and 8, we write L and $\bar{L}_n$ for L^{FY} and $\bar{L}_n^{\mathrm{FY}}$, respectively, since no other loss is considered.

Lemma 7 (Pointwise convergence)**.** *Under Assumptions (A1)–(A6), for each fixed* $v \in \mathbb{S}^{d-1}$,

$$\bar{L}_n^{FY}(v) \to L^{FY}(v) \quad \textit{a.s.}$$

Proof. We follow the same argument as in the proof of Lemma 4. The empirical Fenchel–Young loss can be written as

$$\begin{aligned}\bar{L}_n(v) &= \frac{2}{\gamma n^2} \sup_{y \in \mathcal{P}(n)} \left(\langle s(v), y \rangle - \frac{l_n}{2}\|y\|_2^2 + \frac{l_n}{2}\left\|t^{(n)}\right\|_2^2 - \left\langle s(v), t^{(n)} \right\rangle \right) \\ &= \frac{2}{\gamma n^2} \left(\langle s(v), r_n(s(v)) \rangle - \frac{l_n}{2}\|r_n(s(v))\|_2^2 + \frac{l_n}{2}\left\|t^{(n)}\right\|_2^2 - \left\langle s(v), t^{(n)} \right\rangle \right).\end{aligned}$$

Since S_v has a continuous distribution, the scores $s_1(v), \dots, s_n(v)$ are distinct almost surely. Let π_n denote the corresponding ordering permutation, so that

$$s_{\pi_n(1)}(v) < \cdots < s_{\pi_n(n)}(v) \quad \text{a.s.}$$

For $i = 1, \dots, n$, define $U_{v,i} := F_{S_v}(s_i(v))$. Since $s_{\pi_n(k)}(v) = s^{(n)}_{(k)}(v)$,

$$U_{v,\pi_n(k)} = F_{S_v}\left(s^{(n)}_{(k)}(v)\right).$$

Using $l_n = \gamma/n$ and reindexing the terms according to π_n,

$$\bar{L}_n(v) = \frac{1}{n}\sum_{k=1}^{n} \left(\frac{2}{\gamma} s^{(n)}_{(k)}(v) \left(\frac{1}{n} y^{(n)}_k(v) - \frac{1}{n} t^{(n)}_{\pi_n(k)} \right) - \left(\left(\frac{1}{n} y^{(n)}_k(v) \right)^2 - \left(\frac{1}{n} t^{(n)}_{\pi_n(k)} \right)^2 \right) \right).$$

From Steps 1 and 2 of the proof of Lemma 4,

$$\frac{1}{n}\sum_{k=1}^{n} \left| \frac{1}{n} y^{(n)}_k(v) - U_{v,\pi_n(k)} \right| \le 2K\Delta_n(v), \quad \max_{1 \le i \le n} \left| \frac{t^{(n)}_i}{n} - T_i \right| \to 0 \quad \text{a.s.}$$

Let

$$\tilde{L}_n(v) = \frac{1}{n}\sum_{k=1}^{n} \left(\frac{2}{\gamma} s^{(n)}_{(k)}(v) \left(U_{v,\pi_n(k)} - T_{\pi_n(k)} \right) - \left(U^2_{v,\pi_n(k)} - T^2_{\pi_n(k)} \right) \right).$$

Under Assumption (A6), there exists constant $C > 0$ such that $\|f_i\|_2 \le C$ a.s. for all i, and hence $|s_i(v)| \le C$ a.s. Since $0 \le y^{(n)}_k(v)/n, t^{(n)}_{\pi_n(k)}/n, U_{v,\pi_n(k)}, T_{\pi_n(k)} \le 1$,

$$\begin{aligned}\left| \bar{L}_n(v) - \tilde{L}_n(v) \right| &\le \left(\frac{2C}{\gamma} + 2 \right) \left[\frac{1}{n}\sum_{k=1}^{n} \left| \frac{1}{n} y^{(n)}_k(v) - U_{v,\pi_n(k)} \right| + \frac{1}{n}\sum_{k=1}^{n} \left| \frac{1}{n} t^{(n)}_{\pi_n(k)} - T_{\pi_n(k)} \right| \right] \\ &\to 0 \quad \text{a.s.}\end{aligned}$$

Reindexing the sum gives

$$\tilde{L}_n(v) = \frac{1}{n}\sum_{i=1}^{n} \left(\frac{2}{\gamma} s_i(v) \left(U_{v,i} - T_i \right) - \left(U^2_{v,i} - T^2_i \right) \right).$$

The triples $(s_i(v), U_{v,i}, T_i)$ are i.i.d. with the same joint distribution as

$$(S_v, U_v, T_*) = \left(S_v, F_{S_v}(S_v), F_A(S_{v^*} + \varepsilon) \right).$$

Since the summands are bounded, the strong law of large numbers yields

$$\tilde{L}_n(v) \to \mathbb{E}\left[\frac{2}{\gamma} S_v \left(U_v - T_* \right) - \left(U_v^2 - T_*^2 \right) \right] = L(v) \quad \text{a.s.}$$

Thus, for each fixed $v \in \mathbb{S}^{d-1}$,

$$\bar{L}_n(v) \to L(v) \quad \text{a.s.}$$

□

Lemma 8 (Uniform convergence)**.** *Under Assumptions (A1)–(A6),*

$$\bar{L}^{FY}_n(v) \to L^{FY}(v) \quad \textit{uniformly on } \mathbb{S}^{d-1} \textit{ a.s.}$$

Proof. By Assumption (A6), there exist a constant $C > 0$ and an event Ω_C with $P(\Omega_C) = 1$ such that $\|f_i\|_2 \le C$ for all i on Ω_C. Then, on Ω_C,

$$\|F\|_2 \le \left(\sum_{i=1}^{n} \|f_i\|_2^2\right)^{1/2} \le C\sqrt{n}.$$

Let $c_n(v) := r_n(s(v))/n$ and $d_n := t^{(n)}/n$ as in the proof of Lemma 5. Then

$$\bar{L}_n(v) = \frac{1}{n}\left(\frac{2}{\gamma}\langle s(v), c_n(v) - d_n\rangle - \|c_n(v)\|_2^2 + \|d_n\|_2^2\right).$$

As in the proof of Lemma 5, on Ω_C,

$$\|s(v)\|_2 \le C\sqrt{n}, \quad \|s(v) - s(w)\|_2 \le C\sqrt{n}\|v - w\|_2,$$
$$\|c_n(v)\|_2 \le \sqrt{n}, \quad \|c_n(v) - d_n\|_2 \le \sqrt{n}, \quad \|c_n(v) - c_n(w)\|_2 \le \frac{C}{\sqrt{n}l_n}\|v - w\|_2.$$

For any $v, w \in \mathbb{S}^{d-1}$,

$$\begin{aligned}\left|\bar{L}_n(v) - \bar{L}_n(w)\right| &\le \frac{2}{n\gamma}\|s(v) - s(w)\|_2\|c_n(v) - d_n\|_2 + \frac{2}{n\gamma}\|s(w)\|_2\|c_n(v) - c_n(w)\|_2 \\ &\quad + \frac{1}{n}\left(\|c_n(v)\|_2 + \|c_n(w)\|_2\right)\|c_n(v) - c_n(w)\|_2 \\ &\le \left(\frac{4C}{\gamma} + \frac{2C^2}{\gamma^2}\right)\|v - w\|_2,\end{aligned}$$

since $l_n = \gamma/n$. Let $M = 4C/\gamma + 2C^2/\gamma^2$. Thus, on Ω_C, every $\bar{L}_n$ is M-Lipschitz. Moreover, $\left|\bar{L}_n(v)\right| \le 2C/\gamma + 2$ for all n and v on Ω_C, so $\left\{\bar{L}_n\right\}$ is uniformly bounded.

Applying the same argument as in the proof of Lemma 5 yields

$$\sup_{v \in \mathbb{S}^{d-1}} |\bar{L}_n(v) - L(v)| \to 0 \quad \text{a.s.}$$

□

Lemma 9 (Unique minimizer)**.** *Under Assumptions (A1)–(A6), assume that $\sigma = 0$. Then, $L^{FY}(v)$ is uniquely minimized at v^*.*

Proof. Since $l_n\|y\|_2^2/2 - \langle s(v), y\rangle$ is l_n-strongly convex,

$$\bar{L}_n^{\mathrm{FY}}(v) \ge \frac{2}{\gamma n^2} \times \frac{l_n}{2}\|r_n(s(v)) - t^{(n)}\|_2^2 = L_n^{\mathrm{Sp}}(v).$$

For each fixed v, Lemmas 4 and 7 imply that both sides converge almost surely. Taking limits in the above inequality gives

$$L^{\mathrm{FY}}(v) \ge L^{\mathrm{Sp}}(v).$$

From Lemma 6, $L^{\mathrm{Sp}}(v)$ is uniquely minimized at v^*, and thus $L^{\mathrm{FY}}(v) > 0$ for all $v \ne v^*$. Since $\sigma = 0$, $U_{v^*} = T_*$, and hence $L^{\mathrm{FY}}(v^*) = 0$. Therefore, $L^{\mathrm{FY}}(v)$ is uniquely minimized at v^*. □

Combining Lemmas 7, 8, and 9 and by the consistency of M-estimators (see, e.g., Theorem 2.1 in Newey & McFadden (1994)), we conclude

$$\hat{v}_n^{\mathrm{FY}} \to v^* \quad \text{a.s.}$$

A.4 Details of Example 1

Computational Details First, we verify that the given example satisfies Assumptions (A1)–(A6).

(A1–A2) For any $v = (v_1, v_2) \in \mathbb{S}^1$, $S_v = 2v_1X + v_2Y$ is a linear combination of two independent uniform random variables. Its distribution is continuous and has a density that is positive on the

interior of its support. Hence, F_{S_v} is continuous and strictly increasing on its support. Since the support of S_v is compact, F_{S_v} has a continuous inverse $Q_v(u)$ on $[0,1]$.

(A3) Let $S_{v,1},\dots,S_{v,n}$ be i.i.d. copies of S_v, and define $U_{v,i} = F_{S_v}(S_{v,i})$. By (A1), $U_{v,i} \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(0,1)$. Let $U_{v,(1)} \le \cdots \le U_{v,(n)}$ denote their order statistics. Then,

$$\max_{1\le i\le n}\left|U_{v,(i)} - \frac{i}{n}\right| \to 0 \quad \text{a.s.}$$

Since $s^{(n)}_{(i)}(v) = Q_v(U_{v,(i)})$ and Q_v is uniformly continuous on $[0,1]$ by (A2),

$$\sup_{1\le i\le n}\left|s^{(n)}_{(i)}(v) - Q_v\left(\frac{i}{n}\right)\right| \to 0 \quad \text{a.s.}$$

(A4) For any $v = (v_1, v_2) \in \mathbb{S}^1$, $S_v = 2v_1X + v_2Y$. The two terms are independent uniform random variables supported on intervals of lengths $2|v_1|$ and $|v_2|$. Thus, the maximum of the score density is

$$\|p_v\|_\infty = \frac{1}{\max(2|v_1|, |v_2|)}.$$

Since $v_1^2 + v_2^2 = 1$, we have $\max(2|v_1|, |v_2|) \ge 2/\sqrt{5}$. Thus, for any $v = (v_1, v_2) \in \mathbb{S}^1$,

$$p_v(x) \le \frac{\sqrt{5}}{2}.$$

(A5) From (A4), we may take $K = \sqrt{5}/2$. Hence, we set $\gamma = 1/K = 2/\sqrt{5}$ and $l_n = 2/(\sqrt{5}n)$.

(A6) The support of f is $[0,2]\times[0,1]$, which is compact, and $a = S_{v^*} + \varepsilon = 2X + \varepsilon$, where $\varepsilon \sim N(0,1)$ is independent of f.

We now provide a detailed computation. For the computation below, we let $\varepsilon \sim N(0,\sigma^2)$ and specialize to $\sigma = 1$ at the end.

For $v \in \mathbb{S}^1$, since U_v and T_* are uniformly distributed on $(0,1)$,

$$L(v) = \frac{2}{\gamma}D(v), \quad D(v) := \mathbb{E}[S_vU_v] - \mathbb{E}[S_vT_*].$$

For the true direction $v^* = (1,0)$ and the wrong direction $v' = (0,1)$, $S_{v^*} = 2X$ and $S_{v'} = Y$. Hence,

$$U_{v^*} = F_{2X}(2X) = X, \quad U_{v'} = F_Y(Y) = Y,$$

and

$$T_* = F_A(2X+\varepsilon), \quad A = 2X + \varepsilon.$$

Since Y is independent of (X,ε), $U_{v'}$ is independent of T_*. Therefore,

$$D(v') = \mathbb{E}\left[Y^2\right] - \mathbb{E}[S_{v'}]\mathbb{E}[T_*] = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}.$$

For v^*,

$$D(v^*) = \mathbb{E}\left[2X^2\right] - \mathbb{E}[2XT_*] = \frac{2}{3} - 2\mathbb{E}[XT_*].$$

Let $A' = 2X' + \varepsilon'$ be an independent copy of A. Then, $F_A(2X+\varepsilon) = \mathbb{P}(A' \le 2X + \varepsilon | X, \varepsilon)$. Since $\varepsilon' - \varepsilon \sim N(0, 2\sigma^2)$,

$$\begin{aligned}
\mathbb{E}[XT_*] &= \mathbb{E}[X\mathbf{1}_{\{2X'+\varepsilon' \le 2X+\varepsilon\}}] \\
&= \mathbb{E}\left[X\cdot\Phi\left(\frac{\sqrt{2}(X-X')}{\sigma}\right)\right] \\
&= \int_0^1\int_0^1 x\Phi\left(\frac{\sqrt{2}(x-x')}{\sigma}\right)dx'dx =: I,
\end{aligned}$$

where Φ denotes the standard normal CDF. By symmetry of X and X' and $\Phi(-z) = 1 - \Phi(z)$,

$$\mathbb{E}\left[X' \cdot \Phi\left(\frac{\sqrt{2}(X - X')}{\sigma}\right)\right] = \frac{1}{2} - I.$$

Therefore,

$$2I - \frac{1}{2} = \mathbb{E}\left[(X - X') \cdot \Phi\left(\frac{\sqrt{2}(X - X')}{\sigma}\right)\right].$$

Define $W := X - X'$. Then, its density is

$$f_W(w) = 1 - |w|, \quad -1 \le w \le 1.$$

Therefore,

$$\begin{aligned}
2I - \frac{1}{2} &= \int_{-1}^{1} w(1 - |w|)\Phi\left(\frac{\sqrt{2}w}{\sigma}\right) dw \\
&= \int_{0}^{1} w(1 - w)\left[2\Phi\left(\frac{\sqrt{2}w}{\sigma}\right) - 1\right] dw \\
&= \int_{0}^{1} w(1 - w) \cdot \operatorname{erf}\left(\frac{w}{\sigma}\right) dw \\
&= \left(\frac{1}{6} - \frac{\sigma^2}{4}\right) \operatorname{erf}\left(\frac{1}{\sigma}\right) + \frac{\sigma}{6\sqrt{\pi}} e^{-1/\sigma^2} + \frac{\sigma^3}{3\sqrt{\pi}}(1 - e^{-1/\sigma^2}).
\end{aligned}$$

Since $\mathbb{E}[S_{v^*} T_*] = 2I$, we obtain

$$D(v^*) = \frac{1}{6} - \left(2I - \frac{1}{2}\right).$$

For $\sigma = 1$, $D(v^*) \approx 0.08342 > 1/12$, and thus

$$L(v^*) > L(v').$$

Interpretation of the Population Losses This example can be understood from the different forms of the two population objectives. Since U_v and T_* are uniformly distributed on $(0, 1)$,

$$L^{\mathrm{Sp}}(v) = \mathbb{E}[(U_v - T_*)^2] = \frac{1}{6} - 2\mathrm{Cov}(U_v, T_*).$$

Let $\tilde{S}_v$ be an independent copy of S_v. Then

$$\mathbb{E}[S_v U_v] = \mathbb{E}\left[S_v \mathbf{1}_{\{\tilde{S}_v \le S_v\}}\right] = \frac{1}{2}\mathbb{E}\left[\max(S_v, \tilde{S}_v)\right] = \frac{1}{2}\mathbb{E}[S_v] + \frac{1}{4}\mathbb{E}\left|S_v - \tilde{S}_v\right|,$$

and

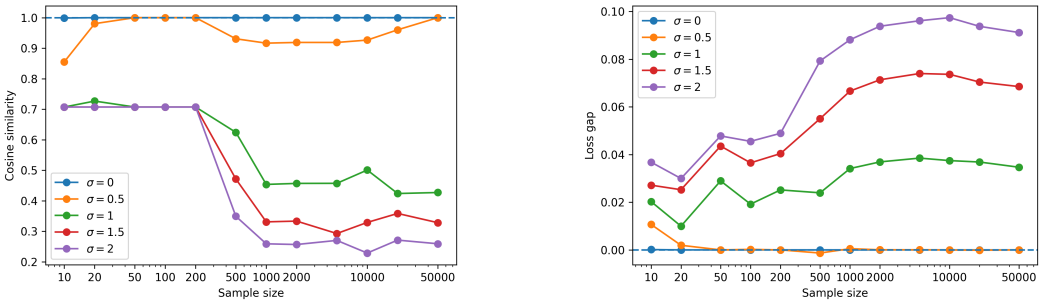
$$\mathbb{E}[S_v T_*] = \frac{1}{2}\mathbb{E}[S_v] + \mathrm{Cov}(S_v, T_*).$$

Thus

$$L^{\mathrm{FY}}(v) = \frac{1}{2\gamma}\mathbb{E}\left|S_v - \tilde{S}_v\right| - \frac{2}{\gamma}\mathrm{Cov}(S_v, T_*).$$

The two objectives use the projected score S_v in different ways. The Spearman loss measures the ordinal alignment between $U_v = F_{S_v}(S_v)$ and T_*. In contrast, the Fenchel–Young loss depends on both the association between S_v and T_* and the dispersion term $\mathbb{E}\left|S_v - \tilde{S}_v\right|$. This dispersion term is smaller for directions with smaller projected score dispersion. Under noise, this term can contribute to an incorrect axis attaining a lower Fenchel–Young loss than the true axis.

Simulation We simulate this example and plot the cosine similarity $\cos(v^*, v_n^{\text{opt}})$ and the empirical loss gap $\bar{L}_n^{\text{FY}}(v^*) - \bar{L}_n^{\text{FY}}(v_n^{\text{opt}})$ where v_n^{opt} is the solution obtained by optimizing the empirical loss on the simulated data.

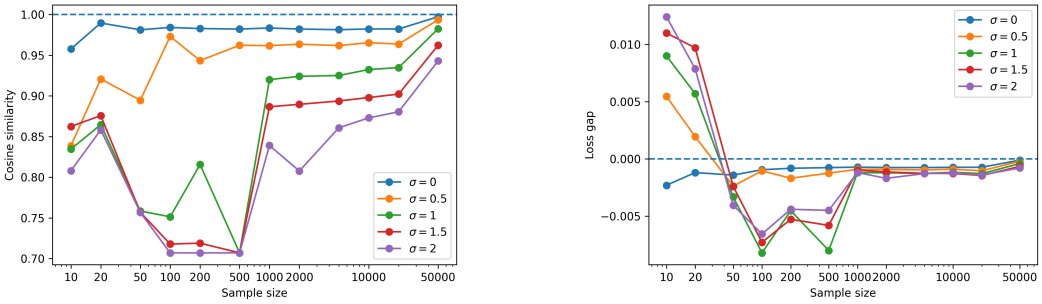


(a) Cosine similarity vs. log sample size

(b) Loss gap vs. log sample size

Figure 4: Cosine similarity and loss gap vs. log sample size for Fenchel–Young loss.

Similarly, we plot the cosine similarity $\cos(v^*, v_n^{\text{opt}})$ and the empirical loss gap $L_n^{\text{Sp}}(v^*) - L_n^{\text{Sp}}(v_n^{\text{opt}})$ for the Spearman loss.



(a) Cosine similarity vs. log sample size

(b) Loss gap vs. log sample size

Figure 5: Cosine similarity and loss gap vs. log sample size for Spearman loss.

Figure 4a shows that consistent behavior for $\sigma = 0$ and 0.5 with cosine similarity approaches to 1 as n increases. In contrast, the cosine similarity does not approach to 1 for $\sigma = 1.0, 1.5$ and 2.0, indicating a failure to recover the true axis. Moreover, Figure 4b shows that the loss gap does not converge to 0 and remains positive for $\sigma = 1.0, 1.5$, and 2.0. A positive loss gap implies that the optimized v_n^{opt} achieves a lower loss value than the true axis v^*. This aligns with our earlier discussion that the true axis may fail to be a minimizer of the Fenchel–Young loss in the presence of noise.

In contrast, the rank axis estimator based on the Spearman loss exhibits behavior consistent with its consistency result in both noise-free and noisy settings. Figure 5a illustrates that $\cos(v^*, v_n^{\text{opt}})$ converges to 1 as n increases for both noise-free and noisy scenario. Moreover, Figure 5b shows loss gap is negative for sample sizes $n \geq 50$ and it converges to 0, which means $L^{\text{Sp}}(v_n^{\text{opt}}) \to L^{\text{Sp}}(v^*)$. These results provide empirical evidence consistent with the consistency of the rank axis estimator based on the Spearman loss.